National Library of Canada
Acquisitions and Bibliographic Services Branch
395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale du Canada
Direction des acquisitions et des services bibliographiques
395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

Canada

UNIVERSITY OF ALBERTA

Strategy Choices in Responding to Multiple Choice Items

by

Ernest N. Skakun ©

A THESIS
SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

Department of Educational Psychology

EDMONTON, ALBERTA

Fall, 1994

National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-315-95265-2

Canada

UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR:	Ernest N. Skakun
TITLE OF THESIS:	Strategy Choices in Responding to Multiple Choice Items
DEGREE:	Doctor of Philosophy
YEAR THIS DEGREE GRANTED:	1994

Permission is hereby granted to the UNIVERSITY OF ALBERTA LIBRARY to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.
The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

Ernest N. Skakun
11716-38A. Avenue
Edmonton, Alberta
T6J 0L9

Date: June 08, 1994.

UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled **Strategy Choices in Responding to Multiple Choice Items** submitted by **Ernest N. Skakun** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy**.

TMaguire
Dr. T.O. Maguire
Supervisor

Charles Harley
Dr. C.H. Harley

S. Hunka
Dr. S.M. Hunka

Rogers
Dr. W.T. Rogers

Lew Romaniuk
Dr. E.W. Romaniuk

G. Page
Dr. G.G. Page
External Examiner

Date: June 08, 1994.

ABSTRACT

Critics of multiple choice items have argued that such items merely test recall and recognition and are not particularly conducive to the assessment of problem solving. Proponents have claimed that such items can measure higher order thinking skills. Very little is known about what transpires when examinees interact with test items and there has been little research on multiple choice items as information processing tasks. The objective of this study was to use a think aloud approach to examine the strategies and reasoning used by medical students in responding to a set of multiple choice items.

The subjects for the study were a convenience sample of 33 third-year and 7 fourth-year medical students who had just completed a twelve-week rotation in medicine. Thirty single-best-answer items from an existing test item library in medicine were selected. The items represented seven different body systems, ranged in difficulty from 19 to 94% and required different tasks.

Students were asked to think aloud while being presented with each multiple choice item. Each interview was audio-taped and later transcribed. The transcripts of the audio-taped protocols were grouped by item and used to obtain an overall impression of how the students approached each item. The notes recorded during each interview, the transcripts and the literature on problem solving, enabled the classification of examinee-item interaction activities or "moves". The moves were grouped into three categories. The first category contained five global item-related moves which dealt primarily with whether examinees generated an answer prior to

reading the alternatives and the actions taken by the examinees with respect to the alternatives. Sixteen moves comprised the second category which dealt with disposition of alternatives. The third category included four activities related to features of successful problem solving.

The list of 25 moves served as a preliminary framework for coding the think aloud protocols of the student-item engagements. Using the list of moves, three broad strategies were detected according to whether or not hypotheses (diagnoses) and answers matching in form the item alternatives were generated prior to reading the list of item alternatives.

The first broad response strategy pertained to a set of items for which students generated few answers matching in form the response alternatives prior to reading the item alternatives. In addition, students did not generate diagnostic hypotheses because this set of items did not request diagnostic hypotheses as a preliminary step to identifying the correct answer. The strategy used most frequently by the students was to read the stem, search the alternatives for an answer, select an answer and eliminate the remaining alternatives by providing a rationale for the incorrect choices.

The second response strategy comes from an item set for which few answers matching in form the item alternatives were generated but diagnostic hypotheses were advanced. For this set of items, students upon reading the stems, activated hypotheses (diagnoses). The generation of diagnostic hypotheses represents an intermediate activity because the items request tasks other than diagnoses. The list of alternatives is searched for an answer and a rationale is provided for the the discarded incorrect choices.

The third response strategy consists of students reading the stem, generating an answer which matches in form the item alternatives, searching the alternatives for a match and discarding the incorrect options.

The strategies illustrate the variations in thought processes encountered. They show progression in sophistication from backward reasoning based on the use of alternatives as provisional hypotheses to the forward, almost script-like expertise in the third response strategy. The claim that multiple choice items assess only recognition can not be sustained in these data. Rather, the reasoning process is influenced by a complex set of constraints involving the semantic and syntactic form of the item, the availability of the appropriate declarative and procedural knowledge organized in efficient structures, and the ability of the examinee to monitor his or her own progress in solving the problem.

The evidence suggests that the problem with multiple choice items is not that they are mere exercises in recognition; rather that it is difficult to predict the processes that will be evoked. There is evidence for reasoning at various levels of expertise. From a psychometric perspective, the next step is to improve the chances of evoking a desired level of expert reasoning by using the think aloud protocols to revise the items. One possibility is to have item writers produce "reasoning scripts" for each item, then compare the intended scripts with the observed scripts. Obvious deficiencies in the item stem and alternatives would become more apparent.

To obtain a better understanding of the nature of multiple choice test-taking behavior, methods such as propositional analysis and structural semantics could be used to analyze the features of the verbal protocols. These methods can most profitably be used after the demand characteristics of the items are better understood. The present research indicates that differences in levels of expertise interact with the nature of the task, the subtleties that distinguish the alternatives and the format, to influence the examinees' reasoning. On this basis, examination of a relatively uniform set of items from the perspective of deeper linguistic analysis would be useful.

Although those investigating the structure of clinical reasoning tend to prefer either propositional analysis or structural semantics, future work clearly needs to integrate the two methods. In reviewing the think aloud protocols, instances were recorded where students used a mixture of elements from propositional analysis (if-then causal rules) and from structural semantic analysis (binary oppositions) to internalize the meaning of information presented in the item. For some items, selection of the correct answer and the elimination of distractors were dealt with primarily from a structural semantic perspective. Hence, use of one method only would be unlikely to capture all that transpires in these student-item engagements.

ACKNOWLEDGEMENTS

I gratefully acknowledge the individuals for their contributions to the completion of this dissertation.

Special thanks are extended to my supervisor, Dr. Thomas O. Maguire, for his guidance, patience, and understanding throughout the duration of the study. The completion of the dissertation means that we can be friends again!

My appreciation is extended to Drs. C.H. Harley and S. M. Hunka, members of the supervisory committee, for their valuable direction. I extend a thank-you to Drs. W.T. Rogers and E.W. Romaniuk, members of the examining committee, for their early comments regarding the theoretical framework for the study. I thank Dr. G.G. Page, external examiner, for this thoughtful and constructive evaluation of the dissertation.

My sincerest appreciation and deepest gratitude is extended to my wife, Vivian, and children Carolynn (a.k.a., Xander), Andrew, and Sheldon for their support and encouragement.

Finally, I acknowledge my parents, Nick and Dorothy Skakun, my mother-in-law, Anne Logozar, and the memory of Metro Logozar, my father-in-law.

TABLE OF CONTENTS

CHAPTER PAGE

I INTRODUCTION 1
The Validity Question 2
Purpose of the Study 9
Organization of Thesis 11

II. SOME RELATED LITERATURE 12
Medical Problem Solving and Expertise 15
Early Studies of Medical Inquiry 16
Expertise in Medicine 23
A Stage Theory of Clinical Reasoning 43
Successful Problem Solving and Expertise 47
Summary 49
Testing and Thinking Aloud 51
Classifying Items 52
Thinking Aloud and Classroom Tests 54
Thinking Aloud and Standardized Tests 58
Identifying Strategies 62
Quality of Thinking 66
Verbalizing and Thinking 70
Summary 72

III. METHODS AND INITIAL ANALYSIS 75
Selection of Examinees 76
Test Items 78
Procedure 83
Initial Analysis 88
Item-Related Activities 91
Disposition of Alternatives 97
Successful Problem Solving Activities 102
Further Analysis 111

IV. RESULTS AND DISCUSSION 115
Responses to the Questionnaire 115
Examples of Think-Aloud Protocols 118
Group One Items 119
Dispensing of Alternatives 132
Strategies for Group One Items 152
Group Two Items 154
Strategies for Group Two Items 189
Group Three Items 191
Strategies for Group Three Items 220
Summary 222

V. DISCUSSION, IMPLICATIONS AND REFLECTIONS ON MULTIPLE CHOICE ITEMS AND CLINICAL PROBLEM SOLVING 223

REFERENCES 232

LIST OF APPENDICES

APPENDIX	DESCRIPTION	PAGE
I	INTERVIEWING MODEL FOR MULTIPLE CHOICE STRATEGIES	244
II	INTRODUCTION TO EXAMINEES	248
III	SAMPLE THINK ALOUD PROTOCOLS	251
IV	PRACTICE AND EXPERIMENTAL ITEMS	255
V	STUDENT QUESTIONNAIRE	269
VI	EXAMPLES OF MULTIPLE CHOICE ITEMS	273
VII	FREQUENCY COUNTS - ACTIVITIES BY ITEM	275
VIII	FREQUENCY COUNTS - ACTIVITIES BY ITEM GROUPS	280

LIST OF TABLES

TABLE	DESCRIPTION	PAGE
1	Associative Strengths of Alternatives	79
2	Distribution of Items by System and Task	80
3	Distribution of Items by System and Difficulty	81
4	Distribution of Items by Task and Difficulty	82
5	Frequencies of Move Use for Items 1 to 5	109
6	Frequency of Move Use for Alternatives of Group One Items	133
7	Frequency of Move Use for Alternatives of Group Two Items	158

LIST OF FIGURES

FIGURE	DESCRIPTION	PAGE
I	Definition of the term "problem" (Smith, 1991)	47
2	Flow chart for Group One items	150
3	Flow chart for Group Two items	188
4	Flow chart for Group Three items	218

I. INTRODUCTION

In today's society, tests play a prominent and critical role. Beside their use in the classroom, tests are widely used in government and industry to make decisions about individual careers (Wainer & Braun, 1988). Tests are also used for entry into various occupations and professions. In the United States, the number of licensed or certified professions and occupations increased from a handful in the early 1900's to about 800 in the early 1980's (Shimberg, 1985). Entry into most professions and occupations requires some form of examination. In Canada, a number of occupations and professions require successful completion of an examination for entry. For example, the Medical Council of Canada has had a licensing process in place since 1912; the Royal College of Physicians and Surgeons of Canada introduced a Fellowship Examination in 1932; and various allied health fields such as Occupational Therapy, Speech Language Pathology and Audiology, Nursing, and Emergency Medical Technicians have embarked on testing programs for assuring the competence of candidates who are seeking entry to the practice of the respective professions and occupations.

In many cases there are two components to the professional testing program, a written component and a practical component. Within the written portion, the multiple choice format has been used because of its potential for broad coverage and simplicity of scoring. Critics of the multiple choice format (for example, Collins, 1990) have argued that such items merely test recall and recognition. Proponents have argued the claim that multiple choice items can measure higher order thinking skills (McGuire, 1987; Feinberg, 1990; Frechtling, 1991; Hambleton & Murphy, 1992).

The purpose of this thesis research is to examine the validity of one component of the examination process, namely, the use of multiple choice items for testing clinical reasoning. This component plays a major role in evaluating the clinical competence of medical practitioners.

The Validity Question

Tests are often administered to permit inferences about some class of behaviors. Such inferences are made on the basis of test scores. More than three decades ago Goodenough (1949) and Loevinger (1957) stated that test behaviors, that is, responses to test items and the test scores which result, are always and essentially *signs* and *samples* of behavior. Test responses as *signs* of behavior represent the presence of traits or underlying processes that may or may not resemble the behavior. Test responses as *samples* of behavior are viewed as representative of some domain of behaviors for which predictions are to be made or inferences are to be drawn. It is obvious that at the heart of any test is the test item. The task for test designers is to generate test items which provide sensitive signs and samples of examinee behaviors. Each item should be demonstrably linked to an appropriate cognitive behavior.

Consider, as an example, the assessment of the behaviors involved in patient evaluation and management by medical students. Test items might be used which require students to gather and interpret information, to generate hypotheses, to priorize hypotheses, to formulate a working diagnosis, and to develop a management plan. The students' responses to these items could be taken as signs of an underlying cognitive process or construct which might be labelled as problem solving or clinical reasoning. In the context of Goodenough's (1949) and Loevinger's (1957) position, the responses are viewed as a sample of the domain of patient evaluation and management. The responses are seen as samples for several reasons. First, among practitioners it is widely accepted that several different approaches may be successful in the evaluation and management of patients (Cutler, 1979; Giannini & Engel, 1986). A student's choice of approach would be a sample. Second, each patient presenting to a clinician, does so with some mix of characteristics such as life span, systems involved, diseases, seriousness, and gender. Each patient is therefore a sample of the many ways in which different patients may present with similar complaints.

Snow and Lohman (1989) and Messick (1989) extended the view of test scores and the responses to test items as signs and samples of behavior. For Snow and Lohman, test performance consists of complex assemblies of information processing activities. Responding to a test item requires that the complex assemblies of knowledge structure be sampled. Hence, from the perspective of cognitive psychology, tests and responses to test items are interpreted as samples of mental organizations.

According to Messick (1989), the current view of validity accepts the realist notion that construct validity lies as the base of all test use. This view, supported by Shepard (1993), acknowledges that it is a construct which "explains performance". Taking the example of evaluation and management of patients, medical educators and researchers talk about the "problem solving" or the "clinical reasoning" construct which is manifested under appropriate conditions. Much effort has been dedicated to devising sets of conditions that will encourage samples or signs of clinical reasoning to occur. For example, multiple choice items, patient management problems (pen and paper or computerized versions), standardized patients, bedside orals, and Cambridge cases based on the concept of key features or competencies (Bordage & Page, 1987) are used to elicit evidence for clinical reasoning. But ultimately, the proof of the validity of the assessment task must be indirect and inferred because the test-taking behaviors that respondents use to arrive at an answer are simply not known. If the construct, clinical reasoning, exists then surely it consists in part of an interesting mix of processing skills, procedural and declarative knowledge components, and strategies or test-taking behaviors.

The multiple choice format has held a prominent position in the technology of standardized testing since 1917 when it was invented and introduced in the Army Alpha and Army Beta group intelligence tests (DuBois, 1970). Critics of the multiple choice item (Collins, 1990; Hoffman, 1962; Muller, 1984; Neufeld, 1985) have noted that the format is not particularly conducive to assessments that call for the application of reasoning to novel problems. Frederiksen (1984) called for more effective item forms that are representative of the curriculum or of instructional objectives, as occurs in the case of achievement tests, or that represent the domain, in the case of standardized tests and tests of competence. Despite this, such tests have been slow to develop. Ward (1985) predicted a decline in the use of the multiple choice item, but, in spite of his claim, the multiple choice item still remains the mainstay in the assessment of medical trainees. It is used at all points in the physician's life from the first year of medical school up to the Fellowship level. Yet the multiple choice item, by the very nature of its rigid format, is left open to influences like guessing, test-wiseness (Rogers & Bateson, 1991), and to considerable debate regarding the constructs assessed (Elstein, 1993; Bennett & Ward, 1993). Previous research (Harvill, 1985; Sarnacki, 1981) has shown that there are some test-taking strategies that are unrelated to the construct being assessed (for example, when in doubt, guess C). Other strategies, such as bringing pieces of information from several sources to argue for the elimination of different alternatives, may be highly relevant to the construct being measured.

While there are many novel formats used in examining medical students, the dominant format is the multiple choice item. Ideally, the task for the writer of multiple choice items is to design items which will elicit the cognitive activities underlying the structural and substantive components of the behavior or construct of interest. Osterlind (1989) and Haladyna and Downing (1989a; 1989b) stated that very little information is available about planning, designing, and writing test items which would reflect significant psychological meaning, although literature abounds on topics such as measurement theory, test construction, and test and item analysis.

This sentiment was echoed by Millman and Greene (1989) and Snow and Lohman (1989). Millman and Greene (1989) perceived that some progress has been made in item writing with the specification of item-writing schemes, but in general they agree with Wesman's (1971) description of item writing as being a creative art.

From the perspective of cognitive psychology, Snow and Lohman (1989) leveled three criticisms at educational psychometric measurement models (i.e., classical test theory, item response theory, and norm- and criterion- referenced measurement). The first criticism is that item performance might have no substantive psychological meaning because item justification is based on how well the item fits the model rather than on some type of cognitive performance. The second criticism is that the measurement models make simple assumptions about the test items such as unidimensionality and item independence. However, it is conceivable that a test score might reflect a combination of strategies, processing skills, and knowledge, and that responses to items might be influenced by previous item responses and contextual effects. Finally, Snow and Lohman stated that the validation of a test's meaning is external to the models and little attempt is made to explain test performance. This last criticism is related to the first; it is unlikely that a test score will have meaning when few, if any, of the items carry any substantive psychological meaning.

In the field of medicine, examinations which use multiple choice items are often criticized because the items tend to assess recall of isolated facts or the recognition of the correct answer from a list of alternatives. Multiple choice items are also criticized by medical faculty because the items fail to assess the goal of medical education, namely, to educate students to be decision-makers, critical thinkers, and problem solvers. In essence, multiple choice items fail to assess how students approach problems and how students apply knowledge. The use of multiple choice items to assess clinical reasoning is suspect (Barrows, 1986; Morgan & Irby, 1978; Muller, 1984; Neufeld, 1985). Eichna (1980), who returned to being a medical student for four years after retiring as a professor of medicine, stated that for most multiple choice examinations only minimal thinking was required and that problem solving was virtually absent. For these reasons, and the fact that most item-writing specifications for multiple choice tests in medicine are inadequate, the construct interpretation of responses from multiple choice items is often suspect.

But are these criticisms of multiple choice items fair? Is it possible that multiple choice items, in addition to assessing recall and recognition, assess at least some components of problem solving or clinical reasoning? Ebel (1979) and McGuire (1987) claimed that multiple choice items can be designed so that more than recall and recognition are assessed. Assuming that there is some interest in this claim, then a question that arises is how evidence relevant to its credibility might be collected. One approach is to use methods such as multimethod-multitrait, exploratory and confirmatory factor analysis, and the analyses of covariance structures. Another approach, advocated by test theorists and cognitive psychologists, calls for examination performance to be supplemented by verbal reports. For example, Cronbach (1971) stated that observing how different examinees respond to test items usually helps to amplify the meaning of the construct. Cronbach goes on to say that in order to get more complete data, examinees should "think aloud" when responding to test items - a view endorsed by Embretson (1983), Anastasi (1988), Messick (1989), and Snow and Lohman (1989).

For Snow and Lohman (1989), one of the ways cognitive psychology can enhance the models of educational psychometric measurement is to analyze existing tests so that the constructs represented by the tests can be better understood. Tests that appear to reflect the cognitive processes involved in, for example, memory, thinking, reasoning, and problem solving can be analyzed using a cognitive information-processing model to gain a better understanding of the mental events and processes that account for responses to test items. In the case of achievement examinations in medicine, the focus would be on the slower and more complex cognitive behaviors where prior knowledge would have an impact on responses to test items. The cognitive information-processing model posits that the behaviors involved in responding to the test items are usually manifested as verbal protocols in which examinees are asked to think aloud as they encounter test items or, alternatively, to describe retrospectively the steps they followed in responding to the items. These verbal reports then serve as the source for discourse analysis. It seems reasonable to seek information from examinees in order to enhance what Snow and

Lohman (1989) refer to as the substantive psychological meaning of test items and the evidence for validity of test scores.

Little is known about what transpires when examinees interact with test items. According to Farr, Pritchard, and Smitten (1990), research focussing on the description of the behaviors and strategies examinees use in taking multiple choice tests is long overdue. More recently, Snow (1993) in discussing the meaning and interpretation of scores derived from different test formats stated that there has been little research on achievement tests as information processing tasks. The goal of the present study is to address this shortcoming.

Purpose of the Study

Recently, studies using verbal reports of examinees responding to test items have been conducted on critical thinking and judgment tasks (Norris, 1990; 1989a; 1989b), strategy choices in addition, subtraction, multiplication, spelling, and balance scales (Siegler, 1989; 1986), time-telling (Siegler & McGilly, 1989), standardized test items (Haney & Scott, 1987), inference ability in reading comprehension (Phillips, 1989), and reading comprehension (Farr, Pritchard, & Smitten, 1990). In medicine, however, the use of think aloud protocols has been restricted to the investigation of what transpires in physicians' minds during the patient-physician encounter (Barrows, Feightner, Neufeld, & Norman, 1978; Elstein, Shulman, & Sprafka, 1978) and to the study of expertise (Kaufman & Patel, 1991; Patel, Groen, & Arocha, 1990; Schmidt, Norman, & Boshuizen, 1990; Patel, Evans, & Groen, 1989; Lesgold, Rubinson, Feltovich, Glaser, Klopfer, & Wang, 1988; Patel & Groen, 1986). A search of the literature failed to identify recently conducted studies which have examined the strategies and the reasoning used by medical students in responding to multiple choice items, despite the great importance of such tests in medical training and licensure.

The first task is to identify and describe the strategies medical students use when encountering multiple choice items. The term "strategies" is used in the broadest sense and refers to a conglomerate of examinee-initiated actions taken by a student when confronted with a test item. "Reading the stem first, generating an answer, searching the alternatives for a match to the generated answer, and eliminating the wrong alternatives" would be an example of a strategy. In this example, the strategy consists of four examinee-initiated activities which are referred to as moves. A combination of moves would constitute a strategy.

The first activity of identifying and describing strategies follows along the lines of the work conducted by Norris (1989a; 1989b), Phillips (1989), and Farr *et al* (1990), except the items are in the field of medicine rather than items dealing with critical thinking, reading inference, or reading comprehension. The second task is to determine whether the strategies associated with answering the items exhibit some of the behaviors of expertise in clinical reasoning, as well as some of the characteristics of successful problem solving.

To carry out these two tasks, forty third- and fourth- year medical students were asked to think aloud as they responded to a set of thirty multiple choice items gleaned from an existing test item library. No attempt was made in this study to address the quality of the thinking or the sophistication of the reasoning process.

Organization of Thesis

The next chapter presents a selected review of the literature on medical problem solving and clinical reasoning, expertise in medicine, and the use of think aloud methods in responding to test items. Chapter III deals with the method and the identification of examinee-initiated activities (moves) from the think aloud protocols obtained from the forty students on the thirty items. In Chapter IV, the moves identified in the initial analysis are examined and discussed in relation to the literature findings on medical expertise and successful problem solving. Examples of students' verbal reports are presented to show the different approaches students used in their interactions with the items. The thesis concludes with a chapter on implications and reflections on multiple choice items and clinical problem solving.

II. SOME RELATED LITERATURE

This chapter is comprised of three sections. Since construct validity lies as the base of all test use and problem solving is at the heart of medical practice, a review of the literature on medical problem solving is presented in the first section. It should be pointed out that researchers used a variety of terms and phrases such as "medical inquiry", "clinical judgment", "decision making", "diagnostic reasoning", "problem solving", and "clinical reasoning" to describe the cognitive activities of the medical students and physicians. Today, problem solving and clinical reasoning are the terms most commonly used (Barrows & Feltovich, 1987; Evans & Patel, 1989; Schmidt, Norman, & Boshuizen, 1990; Smith, 1991).

If the claim that the process exhibited by medical students in responding to multiple choice items is more than just recall and recognition, and may have some elements of problem solving, then it is necessary to have some understanding of the medical problem solving construct. One of the most important early monographs in this area was by Elstein, Shulman, and Sprafka (1978). This monograph introduced the principles of cognitive psychology to medical education and proposed that most of what happened in problem solving could be accounted for by a hypothetico-deductive model. This view was challenged by those who extended the principles of cognitive information processing to the investigation of expertise in medical problem solving. The information processing model led to a delineation of the differences between novice and expert problem solving, and on the basis of the results of several studies, an alternate model of problem solving was proposed by Schmidt, Norman, and Boshuizen (1990). A description of their "stage" model of clinical reasoning is included in the section on Medical Problem Solving and Expertise. In addition to the field of medicine, problem solving has been the focus of study in disciplines such as physics, mathematics, biology, and chemistry. A summary of the characteristics of successful problem solvers (Smith, 1991) and the differences between experts and novices (Holyoak,1991) is also presented in the first section. In addition to the characteristics and behaviors of successful problem solvers, Smith synthesized the results from the various disciplines into a proposed definition of the term "problem". Of particular relevance to the present study is whether the epistemic descriptions of successful problem solving and the expert-novice differences are manifested in the medical students' think aloud protocols.

The second section of Chapter II presents the literature which bears on testing and thinking aloud. It begins with an examination of the degree to which subject matter experts are able to classify test items into the levels of Bloom's (1956) taxonomy or into some modification of the taxonomy. This is important in any discussion of testing and validity because it was thought that the introduction of Bloom's taxonomy would help with understanding the cognitive process involved in responding to test items. Results from several studies showed that the inter-rater agreement for classifying items was not that respectable. Bloom (1978) stated that most of classroom teaching and testing takes place at the lowest level of the taxonomy and as a result, there has been limited understanding of the process used by examinees in responding to test items.

Furthermore, although test designers may have spent considerable time classifying items according to the categories of Bloom's Taxonomy with the hope the item would trigger the desired cognitive process, research evidence for the elicited behavior was either limited or non-existent. Hence, using Bloom's Taxonomy to classify test items was not that helpful in identifying the thought processes used in responding to the items. Several studies which used think aloud methods both with teacher-made examinations and with standardized tests are also reviewed. Most of these studies were aimed at identifying item ambiguities and in doing so, they also identified the strategies and behaviors examinees used in responding to test items. Not only did these studies reveal the imperfections in item design, they also showed that the test-taking behaviors of the examinees often did not match the behaviors intended by the item writers. These studies represent the early attempts to identify the strategies and behaviors exhibited by examinees when encountering test items.

The results of the studies on strategy use and identification contribute to the purpose of this thesis in much the same way as the studies on medical problem solving and expertise. Each adds a facet which enhances the structure and understanding of the construct of problem solving. Because the studies on identifying strategies used a variety of subject areas as the focus of investigation and because examinees ranged from pre-school (four and five year-olds) to college students, the strategies identified in these studies form the basis for developing a list of activities which medical students may initiate in responding to multiple choice items.

In most of the studies dealing with the identification of strategies, attempts were made to relate strategies and reasoning to outcome, that is, selecting or stating an action or answer. The second section of Chapter II also presents a series of studies which investigated the relationship between reasoning and the student's response to a test item. Although the association between quality of reasoning and the actual response is not a priority for the present study, studies of this type may serve as a basis of future research.

One of the arguments against the use of a think aloud procedure is that the act of verbalizing distorts reasoning and performance. The final section of Chapter II presents a study which addresses the concern of distortion. There are other issues of methodology whenever think aloud methods are used and these are addressed in Chapter III.

Medical Problem Solving and Expertise

It was not until the late 1960's and the early 1970's that concerted efforts were made to describe the cognitive activities characterizing the process by which physicians evaluate and manage patients. One of the significant pieces of research on medical problem solving was conducted by Elstein, Shulman, and Sprafka (1978). Their research focussed on process and used the hypothetico-deductive model as the theoretical basis for investigation.

Early Studies of Medical Inquiry

Elstein *et al* (1978) combined think aloud protocols and direct observation of 24 physicians (17 physicians judged criterial and 7 judged noncriterial by their peers) involved in interactions with simulated patients, each with a defined presenting problem. The patient-physician encounters were video-taped and immediately played back to the physician following completion of the interview. During the play-back, each physician was questioned in a non-directive manner as to her/his thinking throughout the interaction with the simulated patient. Elstein *et al* used the three building blocks -- information search units, cues, and critical findings -- to generate 12 different variables which were used in the analysis of verbal protocols. From these results, Elstein *et al* described the process of medical problem solving. Salient features of the process appear in the following description taken from Skakun (1982) .

> At the onset of the physician-patient encounter, the physician perceives a variety of cues from the patient and the environment in which the patient presents. The cues can be observations about the patient, information from the patient's opening remarks, and information that has come with the patient. The set of verbal and non-verbal cues that the physician can select from all the information available is largely determined by the physician's clinical experience and the setting in which the encounter takes place. The cues represent the initial data about the patient's problem and are assembled by the physician into an initial concept of the problem. It is this initial concept which starts the whole direction and scope of the clinical reasoning process.

Almost simultaneously with the development of the initial concept, the physician generates from two to five hypotheses as possible explanations for the patient's problem. The term hypotheses refers to ideas, hunches, guesses, impressions, or even diagnoses. The hypotheses may be broad and vague or they may be focused and specific and serve as a guide for the physician's interview and examination of the patient. All the hypotheses that the physician will entertain throughout the encounter are usually developed within the first quarter of the physician-patient encounter. Throughout the encounter, the questions that are asked and the physical examination that is performed are chosen to confirm or refute, or strengthen or weaken the likelihood of the generated hypotheses. The hypotheses are processed in a parallel rather than a sequential manner.

The questions a physician asks to establish, shape, refine, strengthen, or refute the hypotheses are search questions. In addition, the questions are chosen to produce data that relate to at least two hypotheses. In this manner, hypotheses are verified, and if verification is difficult, they are ranked in order of likelihood. When questioning yields data that lead to a blind alley where confirmation or ranking is no longer possible, the questions asked are of a scanning type. These are questions that seek new cues or data that might change the initial concept of the patient's problem. Scan questions take the form of functional inquiries or review of organ systems and questions about the patient's background and previous medical history. These questions play a minor role in the clinical reasoning process; their major functions are to look for new cues that might indicate some aspect of the problem that was overlooked, fill in background data, satisfy the needs of the more compulsive physician and increase the confidence that the decisions made about the patient's

> problem are correct (Barrows & Tamblyn, 1980). Thus in resolving the patient's problem, the physician employs a problem-oriented search strategy which is disciplined and logical.
>
> The physical examination is conducted in a search mode to confirm any hypotheses that still remain likely after the completion of the history. Barrows and Tamblyn (1980) state that the physician knows exactly what he is looking for in the physical examination. Laboratory tests and special investigative procedures also form part of the clinical reasoning process and are employed by the physician to help sort out hypotheses.
>
> The majority of the patient-physician encounter is taken up with history taking and physical examination. In this activity, the physician selects a strategic sequence of inquiries related to the hypotheses that he has generated. The results of the inquiries help the physician confirm or refute or rank the hypotheses. At some stage in the encounter the physician decides that he has obtained all the information that is available and that it is time to reach closure on the patient's problem. At closure, several options are available to the physician. He may choose to do nothing, await the results of some tests and then treat and manage the patient, or he may decide to intervene with management and treatment, or hospitalization. If he decides to treat the patient, he may do so medically, surgically, pharmacologically, or psychologically. (Skakun, 1982, p. 733).

In summary, the model of clinical reasoning proposed by Elstein *et al* (1978) encompassed the four major activities of data acquisition, data interpretation, hypotheses generation, and hypotheses evaluation. Because the physicians' thinking combined early hypothesis generation with deductive inference, the model was labelled by Elstein and his co-workers as hypothetico-deductive. The hypothetico-deductive method represented an alternative to the two other types of scientific reasoning -- inductive and deductive. To what extent the method is deductive is debatable. Ridderikhoff (1989) claims the hypothetico-deductive method to be inductive while Anzai (1991) states that the work of Elstein and his colleagues suggests physicians tend to generate several hypotheses on the basis of recognized relevant information (forward reasoning or induction) and then additional data are gathered to support the generated hypotheses (backward reasoning and deduction).

Considering the state-of-the-art with respect to the influence of cognitive psychology in medical education at the time Elstein *et al* (1978) and Barrows and his colleagues (Barrows *et al*, 1978) conducted their respective projects, their findings on clinical reasoning showed a departure from the conventional manner in which medical students were taught and expected to reason about patient problems. In their very early encounters with patients, medical students were taught to take a complete history, do a physical examination, and order the routine laboratory tests. Data gathering is generally unfocussed. Armed with this wealth of data, students are then asked to make order of the collected data and to arrive at a diagnosis. Diagnoses are generated from the acquired evidence as well as from memory recall. Some of the generated diagnoses may have very little or no relation to the patient's problem. Each of the diagnoses is falsified on the basis of the collected data. The process continues until the student is left with a single diagnosis. According to Ridderikhoff (1989; 1991), this approach to teaching the resolution of patient problems is largely deductive and may not necessarily represent the way successful problem solvers think. Deductive reasoning means going from theory to facts (from hypotheses to data), where a problem is resolved by generating new concepts to account for existing data. Reasoning from hypotheses to data (backward chaining, means-end analysis) may be used by physicians who are working on problems outside their area of expertise. This is in contrast to the thinking exhibited by experts and successful problem solvers who, when faced with a problem, use inductive and forward reasoning (forward chaining) among other strategies to arrive at an answer. In medicine, forward reasoning is evidenced when physicians generate their diagnoses by studying the patients' signs and symptoms, that is, hypotheses are generated on the basis of the information obtained during the various stages of the physician-patient encounter and explanations of the problem are based on the acquired evidence. Forward chaining is characterized by reasoning which moves from the given information (for example, signs and symptoms) to the unknown (diagnosis). In contrast, backward reasoning is witnessed when physicians work from the diagnosis, that is, a hypothesis of the unknown, back to the given information (Patel & Groen, 1991; Ericsson & Smith, 1991). In such cases, the resolution of the problem may be inaccurate or incomplete. Highly successful problem solvers and experts, working in their domain, resolve problems by reasoning in a forward fashion.

A study by Barrows, Feightner, Neufeld, and Norman (1978) paralleled the Elstein *et al* (1978) study. Using standardized patients to simulate patient problems, Barrows and his colleagues administered four different problems to a sample of medical students, Family Physicians, and Internists. Sample sizes vary for each problem. For example, the number of students varies between 13 and 17 across the four problems. The total number of student encounters is 63 while the number of encounters for Family Physicians is 30 and the number of Internist encounters is 30. With respect to the results, Barrows *et al* (1978) concluded that the clinical encounter is characterized by the formation of multiple hypotheses very early in the encounter and that these hypotheses play a central role in the subsequent search for information to support the generated hypotheses. The use of hypothetico-deductive reasoning during the patient encounter by both students and clinicians was confirmed by Barrows *et al* (1978).

The studies conducted by Elstein *et al* (1978) and Barrows *et al* (1978), which were rooted in the information processing approach developed by Newell and Simon (1972), had several aims. First, they attempted to structure the theoretical basis of discovery learning, which had been introduced to teaching and learning in the 1950's, but which was not very successful. Second, these studies tried to relate the clinical reasoning skills of experienced physicians to a set of strategies and heuristics which were largely domain-independent and used in general problem solving (Elstein, Shulman, & Sprafka, 1990; Groen & Patel, 1985; Schmidt, Norman, & Boshuizen, 1990). Third, these studies focussed on general parameters of performance, such as number of hypotheses generated, number of cues acquired, and accuracy of interpretation rather than on the organization and the processing of knowledge specific to the content of the patient's problem (Kaufman & Patel, 1991).

At least three unexpected findings emerged from the work conducted by Elstein, Barrows, and their respective colleagues. The first was content specificity. Performance on one case or patient problem had little relation to performance on another problem and performance was highly dependent upon knowledge relevant to the resolution of the specific patient problem. Second, there were no differences in performance between criterial and noncriterial physicians and between medical students and physicians. The expectation was that individuals with more education and/or experience would gather more relevant data and therefore score higher. Finally, thoroughness of data collection was unrelated to accuracy of data interpretation.

To appreciate some of these unexpected findings it is necessary to understand that at the time (late 1960's and early 1970's) Elstein *et al* (1978) and Barrows *et al* (1978) conducted their studies of clinical reasoning, research in cognitive psychology was primarily concerned with the first phase of the Newell and Simon (1972) theory of information processing. Referred to as the first generation theories of expertise by Holyoak (1991), problem solving was conceptualized as consisting of a small number of search heuristics that could be applied across a broad range of domains. It was not until after the publication of the results from the Elstein and Barrows studies, that research in medical problem solving moved into the second generation which considered the influence and importance of domain-specific knowledge in problem solving.

Elstein's work dealt with the first phase and was an application of the de Groot (1965) and Chase and Simon (1973) studies on general problem solving abilities of grand master and novice chess players to the field of medicine. The results of the chess studies showed that grand masters and novices did equally well in recalling random layouts of board pieces. On layouts which emerged in the natural course of a game, grand masters recalled more of the positions than did novices. These results suggested that grand masters were using knowledge which was specific to chess and that prior knowledge and experience affected problem solving (Perkins & Salomon, 1989). This finding from the chess studies coupled with the result from Elstein *et al* (1978) of poor differentiation between better and weaker clinicians led to a shift in research on medical problem solving. Studies on the process of solving problems gave way to research which focussed on the organization and availability of domain-specific knowledge required to reach a diagnosis. To obtain a better understanding of the structure of clinical reasoning, researchers began investigating differences in thinking between novices and experts.

Expertise in Medicine

The studies in chess along with the unexpected findings in medical problem solving which showed both case specificity and the similarity of performance between novices and experts, gave rise to research which shifted from investigating domain-independent general problem solving to problem solving which was contextual. The use of the hypothetico-deductive approach by both novices and experts was being questioned (Groen & Patel, 1985), and a greater emphasis was placed on the content, nature, and the use of knowledge required in resolving a problem (Kaufman & Patel, 1991). In addition, to understand in more detail the problem solving construct, research moved to the investigation of differences between novices and experts. This move was also partly motivated by the research findings on novice-expert differences in fields such as biology (Smith,1991), chemistry (Bodner, 1991), physics (Chi, Feltovich, & Glaser, 1981; Schultz & Lochhead, 1991), and mathematics (Greeno, 1991).

Several studies in medicine modelled after the novice-expert chess player studies have been conducted. Muzzin, Norman, Feightner, Tugwell, and Guyatt (1983) investigated whether differences in recall of clinical scenarios between experts and novices would be enhanced if exposure to the scenarios was brief. This study was a follow-up of the Norman, Jacoby, Feightner, and Campbell (1979) and the Muzzin, Norman, Jacoby, Feightner, Tugwell, and Guyatt (1982) studies which had beginning second year medical students and practicing internists recall information from case protocols. Norman and colleagues used "typical" and "random" protocols whereas Muzzin *et al* used "typical" and "atypical" protocols. Participants were given two minutes to read each case after which they were asked to free-recall the information presented in each scenario. Norman *et al* (1979) reported differences between beginning second year medical students and practicing internists in recalling typical case presentations and no differences in recall of the random case descriptions. Experts showed better recall of clinically-relevant diagnostic information. The study was unable to demonstrate any difference between experts and novices in the actual number of items recalled.

The Muzzin *et al* (1982) study was similar in format except that "random" cases were replaced by "atypical" cases and the testing time was standardized. Instead of having the participant read the case protocol for two minutes, the participant listened to a tape-recording of the case presentation and then read a written transcript of the case. At the end of the reading, each participant was asked to free-recall items of information about the case and then to offer a differential diagnosis. Eight cases (four typical and four atypical) were presented in random to eight consulting internists (experts) and to eight Year 2 students (novices). In this study, no differences between experts and novices were found in the number of items recalled for either typical or atypical cases. However, the average number of items recalled per "run" (where a "run" is defined as verbal recall framed by a minimum of 2 seconds), showed a difference between experts and novices for all cases. The "runs" of experts contained about three times as many items of information as those of novices. Several other findings emerged from this study. Experts tended to report information in terms of coherent patterns whereas novices viewed the information in isolated fragments. Second, different information was recalled. Novices recalled some of the more clinically irrelevant material and more of the negative findings. Experts tended to recall the more important and clinically relevant information. Third, both novices and experts took less time to recall the typical cases. Finally, experts generated diagnostic hypotheses almost immediately into the case whereas novices generated hypotheses as an afterthought.

Five rheumatologists (experts), six subspecialists (experts) in cardiorespiratory disorders, 10 residents in internal medicine, and 10 beginning third year medical students (novices) served as subjects for the study conducted by Muzzin *et al* in 1983. As mentioned earlier, the purpose of their study was to determine whether expert-novice differences in recall would be enhanced if exposure to cases was limited. A total of 16 protocols of cases in the areas of rheumatology and cardiorespiratory were developed. The subspecialists were given 45 seconds to read each approximately half-page protocol in their specialty after which they recalled information presented in the protocol. The same procedure was followed with the residents and students with the following modification. Half of the protocols were presented for 45 seconds and half were presented for two minutes. Combining the results from this study and the one conducted in 1982, Muzzin *et al* (1983) reported no differences in the mean number of items recalled between students, residents, and specialists for two minute exposures to the cases. For the 45-second exposure time, subspecialists recalled fewer items than novices for the cardiorespiratory cases; no differences were noted for the rheumatology cases. In addition, novices' recall tended to be verbatim and sequential while experts tended to select key features of the cases and to group these features into a meaningful category or disease condition.

Neither Norman nor Muzzin provided examples of typical, atypical, or random cases, and therefore some difficulties arise in generalizing these results. Several succeeding studies, however, provided either descriptions or examples of the cases. In a study conducted by Claessen and Boshuizen (1985), typical and atypical variants were based on real patient records. In the typical cases, all items had values typically found in patients presenting with the problem. In the atypical cases, the values of some of the items were changed to make the items no longer typical of the disease. However, the probability of the intended diagnosis was not reduced with this change. Case descriptions were presented to second (N=7), fourth (N=8), and fifth (N=6) year medical students and to six physicians. A second sample of ten third-year students, nine fourth-year, and seven fifth-year medical students from another medical school was also used. Claessen and Boshuizen found that the number of errors in recalling material did not decrease with experience and that the number of errors generated for the typical and atypical protocols were about the same. Unlike the results of Norman *et al* (1979), typical cases were not recalled any better than atypical cases. Experts did not recall the cases better than students. Claessen and Boshuizen (1985) also found that with increasing experience and education, participants increasingly grouped protocol elements in a way which was compatible with the illness script. In contrast to Muzzin *et al* (1982), Claessen and Boshuizen did not find evidence for diagnostic hypotheses facilitating recall.

Attempts to reproduce the results with chess masters and novices obtained by de Groot (1965) and Chase and Simon (1973) in medicine using "typical" and "atypical" medical scenarios produced mixed results. Patel, Groen, and Frederiksen (1986), using propositional analysis, re-analyzed the recall protocols of six students and six physicians for one typical and one atypical case from the Muzzin *et al* (1982) study. Rather than considering the number of items of information recalled, Patel *et al* generated items of recall and items of inference. Recall was any item in the subject's protocol that corresponded exactly to the information presented in the original text. Information that was transformed by the subject was referred to as an inference. In addition to the identification of recall and inference items, propositions were classified as disease-relevant or disease-irrelevant. Physicians showed minimal differences on the two cases in terms of the number of relevant propositions they recalled and inferred. Students recalled more than they inferred. On the atypical case, both students and physicians inferred more propositions than they recalled. With respect to the disease-relevant propositions, physician performance was similar on the two cases in terms of the number of propositions they recalled and inferred. Students recalled and inferred more relevant propositions for the typical case than for the atypical case. Physicians and students showed differences between typical and atypical cases in terms of the number of irrelevant propositions recalled and inferred. In summary, these results showed that given cases that are similar to those seen in practice, experts make more inferences on highly relevant information than do novices. Students recall and infer more of the low relevance information. Thus, using a different method of analysis, Patel, Groen, and Frederiksen (1986) were able to demonstrate differences in recall and inference between novices and experts.

Because the use of atypical cases in medicine produced unclear results, Coughlin (1986, cited in Groen & Patel,1988) modified the research model by adding a third component in which the text making up the case was scrambled. In the scrambled version, items of patient description, complaint, history, physical examination, and laboratory findings were intermingled with each other. For the scrambled routine cases, there were no differences between experts and novices. In the unscrambled versions, while experts and novices recalled about the same amount of irrelevant information, experts tended to make more inferences. In addition, all the experts and most of the novices reorganized recalled text in the same way in which physicians summarize patient information. These are ordered categories beginning with a description of patient, patient's presenting complaints, information obtained from the history, physical examination findings, results from laboratory tests if such tests are requested, differential diagnoses, working diagnosis, and management.

The conclusion from the Elstein *et al* (1978) and Barrows *et al* (1978) studies that experts and novices generate and test hypotheses in the same manner was inconsistent with the findings reported in domains other than medicine. This led Patel and her associates to conduct a number of studies on problem solving strategies. In many of these studies the same clinical case or cases were used and depending upon the research question, some variant of the following paradigm served as a basis for investigation. Subjects were first presented with a written description of a clinical case. They were given a fixed amount of time (usually 2.5 minutes) to read the case after which the description was removed. Two pieces of information were then requested from the participants -- a written free-recall protocol of the case followed by a written explanation of the underlying pathophysiology of the case. Subjects had access to neither the case nor the free-recall protocol when the pathophysiology was being explained. Finally, subjects were asked to provide a diagnosis.

Propositional analysis techniques (Frederiksen, 1975; Kintsch & van Dijk, 1978) were then applied in which the propositions in the written clinical case and in the recall protocol were mapped unto the propositions in the pathophysiology protocol. The recall frame probes comprehension while the pathophysiology frame considers the subjects' understanding of the causal networks that explain the causes and consequences of diseases. Since there is a long tradition in medicine of describing problems in terms of cause-effect relationships, propositional analysis provides a means of examining the relationship between propositions and the semantic network representations of the explanations provided by the subjects.

Using the above framework, Patel and Groen (1986) investigated several aspects of the problem solving process in novices and experts. One of these dealt with the use of forward and backward chaining in reasoning. Forward reasoning occurs when a series of clinical-causal rules are produced and used to move the reasoning from uncertainty (patient data) to certainty (diagnosis or explanation of patient data). New facts are deduced from existing data and all newly generated data by the physician is based on the original facts. In contrast, backward reasoning is characterized by introducing new facts to account for the original data. More information is added to the problem space without necessarily establishing any causal connectives between the newly generated information and the existing facts. In backward reasoning, the physician works from a hypothesis regarding the patient problem back to the given information and checks to see whether the information fits with the hypothesis (Ericsson & Smith, 1991; Patel & Groen, 1991).

In investigating directionality of reasoning, Patel and Groen (1986) had seven cardiologists read the following written case of a patient with bacterial endocarditis.

> This 27 year old unemployed male was admitted to the emergency room with the complaint of shaking chills and fever of four days' duration. He took his own temperature and it was recorded at 40 C on the morning of his admission. The fever and chills were accompanied by sweating and a feeling of prostration. He also complained of some shortness of breath when he tried to climb the two flights of stairs to his apartment. Functional inquiry revealed a transient loss of vision in his right eye which lasted approximately 45 s on the day before his admission to the emergency ward. Physical examination revealed a toxic looking young man who was having rigor, his temperature was 41 C, pulse 120, BP 110/40. Mucus membranes were pink. Examination of his limbs showed a

> puncture wound in his left antecubital fossa. The patient volunteered that he had been bitten by a cat at a friend's house about a week before admission. There were no other skin findings. Examination of the cardiovascular system showed no jugular venous distention, pulse was 120 per minute, regular, equal, and synchronous. The pulse was also noted to be collapsing. The apex beat was not displaced. Auscultation of his heart revealed a 2/6 early diastolic murmur in the aortic area and funduscopy revealed a flame shaped hemorrhage in the left eye. There was no splenomegaly. Urinalysis showed numerous red cells but there were no red cell casts. (Patel and Groen, 1986, p. 97).

This case has four main components which are necessary to generate a complete and accurate diagnosis of acute bacterial endocarditis. These key features are infection, aortic valve insufficiency, emboli, and acuteness of the presentation. An example of forward reasoning appears in the following pathophysiological protocol of a cardiologist who gave the correct diagnosis.

> The important points are the acute onset of chills and fever in a young male with puncture wounds in the left antecubital fossa indicating high probability of drug use and therefore susceptible to endocarditis. The history of transient blindness in one eye supports an embolic phenomena for a left valvular vegitation. The shortness of breath (SOB) on exertion and the early diastolic murmur plus wide pulse pressure support aortic insufficiency and thus aortic valve endocarditis. The normal spleen size indicates a more acute process. The urinary findings support renal emboli. The history of being scratched by a cat raises the differential diagnosis

> of cat scratch fever, but there are no other supporting findings. (Patel and Groen, 1986, p. 99).

In the above protocol the physician links critical information such as "young, unemployed, puncture wounds" with "intravenous drug use" which combined with the "embolic phenomena" generates the hypothesis of bacterial endocarditis. This protocol can be contrasted with the following protocol in which backward reasoning leads to an incorrect diagnosis.

> General bacteremia and septicaemia are secondary to bacteremia resulting in CNS response to septicemia (chills, fever) and with seeding of organisms on the aortic valve resulting in subacute aortic insufficiency due to leaflet destruction and hence a collapsing pulse but without sufficient severity or duration to cause clinical heart failure or cardiomegaly. Possible emboli from the valve. Transient blindness in left eye and flame hemorrhage in the other. (Patel and Groen, 1986, pp.114-5).

According to Patel and Groen, the physician in the above protocol began with a general hypothesis about bacteremia and worked backward to the facts presented in the case.

Because the patient contacted the infection through unusual means (contaminated needle from intravenous drug use), this made the case atypical. Of the seven cardiologists, four gave the correct diagnosis and all four used forward reasoning in arriving at the diagnosis. The remaining three physicians did not make use of the critical information but instead made use of irrelevant rules and backward reasoning. When the response protocols of the written case were analyzed using propositional analysis, Patel and Groen were unable to detect any differences between physicians with accurate and those with inaccurate diagnoses with respect to the recall of relevant or irrelevant propositions.

To determine whether forward and backward reasoning is a function of relevant prior knowledge, Patel, Groen, and Arocha (1990) in the first of two experiments, compared the method of reasoning of the seven cardiologists used in the 1986 study with the performance of six surgeons and six psychiatrists. The same case of acute bacterial endocarditis was used. Only one of the surgeons gave an accurate diagnosis and this was reached by pure forward reasoning. Psychiatrists gave either partially correct or incorrect diagnoses. Forward reasoning was not demonstrated by any of the participants who gave inaccurate or incomplete diagnoses.

With respect to the recall of propositions in each of the four main components of infection, emboli, aortic valve insufficiency, and acuteness, there were no differences between cardiologists, surgeons, and psychiatrists. Likewise, there were no differences in the recall of relevant propositions between those making accurate, partially accurate, or inaccurate diagnoses. However, differences were noted in the number of rules used in the pathophysiological explanation. Physicians making accurate diagnoses used more rules than physicians making partially correct diagnoses. Those with inaccurate diagnoses used the least number of rules. Patel, Groen, and Arocha (1990) concluded that performance on the acute bacterial endocarditis case was determined by whether or not the physicians possessed the relevant knowledge. In addition, forward reasoning was a necessary and sufficient condition for arriving at an accurate diagnosis.

In a second experiment, Patel *et al.* (1990) investigated whether the relationship between forward reasoning and diagnostic accuracy would be affected when physicians were presented with more ill-structured cases. Two written cases - one on *Hashimoto's thyroiditis* and the second on *cardiac tamponade with pleural effusion* were presented to eight cardiologists and eight endocrinologists. In each group of specialists, half were researchers and half were clinical practitioners. For the thyroid case, all four of the endocrine researchers and three of the four endocrine clinical practitioners made the correct diagnosis; all eight endocrinologists made inaccurate diagnoses on the cardiology case. On the cardiology case, two of the four cardiology researchers and three of the four clinical cardiology practitioners made accurate diagnoses. For the endocrine problem, seven of the eight cardiologists provided incomplete diagnoses. In general researchers used a pathophysiological basis for reaching a diagnosis while the practitioners used the clinical features of the case.

Similar to the findings reported from the first experiment, cardiologists and endocrinologists were able to recall the relevant information in the text and provide a summary of the key features. Patel *et al* were also interested in the relationship between the recall and the pathophysiology protocols. Propositions in the recall protocol were matched with the explanations provided in the pathophysiology protocol. Researchers working in their own specialty repeated about half of the recall protocol in the pathophysiology protocol. In solving the case outside their specialty there were fewer repetitions between the two protocols. For practitioners, the pattern was reversed. For the endocrine problem, 14% of the propositions overlapped and only 7% overlapped on the cardiology case. On cases outside their domain, practitioners used more information from the recall protocol to explain the pathophysiological underpinnings of the case.

Several reasons may account for these differences between researchers and practitioners on specialty and non-specialty problems. For the specialty-specific problems, it is conceivable that the four cardiologists and four endocrinologists who spent the majority of their time in research had limited contact with patients and clinical practice. Their knowledge base with respect to disease processes would be predominantly pathophysiological and based on textbook descriptions of disease presentation. To make meaning of the case presentation, researchers' explanations would consist of an overlap of facts between the recall and the pathophysiological protocols. The recall and the explanation of the case may be a reflection of the researchers' attempts to integrate the facts. In contrast, clinicians would have replaced the classical textbook presentations of disease pathophysiology with a pathophysiology which would be an integration of the repertoire of patients presenting with the illness and which the practitioners had managed.

For cases outside their specialty, the overlap between recall and pathophysiology protocols was less than the overlap between the two protocols for the practitioners. Researchers may have viewed the recall and explanatory tasks as separate activities where integration of facts may have proved to be difficult whereas the practitioners may have attempted to integrate the information presented in the case with the classical presentation of the illness.

Patel *et al.* (1990) also looked at patterns of reasoning. Forward reasoning characterized the thinking patterns of those practicing clinicians generating accurate diagnoses. These same individuals used backward reasoning but only to clarify one or two pieces of information. For problems outside their specialty, practitioners used a mixture of forward and backward reasoning. For problems in their specialty, researchers displayed the same pattern of reasoning as the practitioners. However, with problems outside their domain, reasoning was predominantly backward and usually led to inaccurate or incomplete diagnoses. These findings suggest that forward reasoning occurs in the presence of adequate domain knowledge and backward reasoning is reverted to when knowledge is deficient.

For most of the early investigations of forward and backward reasoning, Patel and her colleagues used subjects who were qualified physicians or researchers. Ramsden, Whelan, and Cooper (1989) investigated the approaches used by fifty fourth-year medical students to understand the information presented in the case histories of two patients. After each student had read and thought about each written patient case, a partially structured interview was used to elicit information from the student about the case. Students were asked to provide and justify diagnoses. The authors noted two broad approaches in the students' diagnostic problem solving. The first of these, referred to as ordering, is characterized by isolation and fragmentation. Students work with single or small groupings of findings, make short associative links, and may formulate diagnoses. However, all the information presented about the patient is not integrated and accounted for in a meaningful way. The more sophisticated approach, referred to as structured, is characterized by completion and structure. Multiple links are made between groupings of information or between clusters of patient signs and symptoms, the underlying pathophysiology, causes of the patient's problems, and the diagnoses.

Ramsden *et al* (1989) identified four different diagnostic strategies used by the students. In the ordering approach, the more sophisticated strategy was that of exclusion. Students selected a piece of patient information and then explored it. A diagnosis would be selected by associating it with a specific piece of data (for example, lung carcinoma would be related to smoking). Diagnoses would be ruled out because of the absence of other related and necessary clinical features. The ruling out process would continue until the student was left with a single diagnosis. All remaining patient information which had not been discarded with the other diagnoses would be explained on the basis of the one last remaining diagnosis. The second strategy used by the students was pattern matching. A diagnosis would be offered because the diagnosis was associated with one or more bits of patient information. The remaining patient information would be either ignored or believed to fit the diagnosis. Both exclusion and pattern matching strategies have some characteristics in common with backward reasoning.

In the structured approach, which resembles more of a pure forward reasoning way of dealing with problems, diagnoses were activated using either a stepwise pathophysiological or a diagnostic integration strategy. Students using the pathophysiological strategy moved from patient information to diagnosis in a forward fashion using pathophysiology as a framework for explaining the patient's clinical features. Pathophysiology also served as the undergirding mechanism for the diagnostic integration strategy. The difference between the two strategies in the structured approach is that with diagnostic integration, students were able to explain clusters of clinical findings in parallel with other clusters of signs and symptoms and the relation of these clusters to the activated diagnoses. The rule out and rule in technique for dealing with diagnoses was also in use in the structured approach. Evidence for either accepting or rejecting a diagnosis was supported by the presence or absence of clinical information. The final or working diagnosis selected by the student was one which accounted for most or all of the patient information presented in the case.

Ramsden *et al* (1989) conclude that the students showed a great deal of variability in dealing with the two written cases. The majority of students used the ordering approach and therefore the exclusion and pattern matching strategies for dealing with the patient information. Both approaches have elements of forward and backward reasoning although the structured approach is more characteristic of pure forward reasoning.

From the studies on propositional analysis and using novices, experts, and students as subjects, several findings emerge which embellish the problem solving construct. Groen and Patel (1991) report that differences between novices and experts in free recall of material exist and are consistent with the findings obtained by Chase and Simon (1973) in their study of chess players. Experts recall more of the material relevant to a diagnosis and filtered out irrelevant material. Experts comprehend and integrate new information into their existing knowledge base.

Patel, Evans, and Groen (1989) summarize some of the other characteristics of experts and novices. For the purposes of simple recall, novices and experts have the ability to comprehend case descriptions. Experts can solve diagnostic problems. Novices can not solve diagnostic problems; their diagnostic accuracy, however, increases with exposure to clinical information.

Experts are able to maintain both local and global coherence in their diagnostic protocols. Local coherence refers to the use of characterization, restatement, and questioning, whereas global coherence consists of summarization and explanation. For novices, local coherence improves with additional basic science knowledge and global coherence improves with exposure to clinical situations. Experts tend to explain a patient case in a clinically-oriented manner whereas novices incorporate pathophysiological data and beginners use circumstantial features to explain the patient's problem. Experts are more selective and discriminatory in the use of patient data and report fewer but highly relevant findings. Novices tend to use more findings. Novices tend to rely on the basic sciences for their explanations whereas the experts focus on clinical situations. Because experts are typically individuals with more education and experience than novices, it is not too surprising to find experts relying on clinically-based models for their explanations and inferences. These behaviors are related to the activities that dominate an expert's livelihood. Novices, on the other hand, generally have not completed their formal undergraduate education, their knowledge base is predominantly pathophysiological in nature, and their contact with patients is limited.

Another difference between experts and novices relates to the type of reasoning which is used. Experts operating within their domains tend to reason in a forward manner (generate hypotheses and use data on the basis of the information presented in the case) whereas novices tend to reason backward (seek data on the basis of hypotheses and generate data on the basis of information not necessarily presented in the case). Experts operating outside their domain revert to backward reasoning or use a mix of the two types of reasoning. Forward reasoning is usually associated with successful performance while backward reasoning emerges when knowledge is inadequate to resolve the problem.

Most of the literature on medical expertise is referenced to the work of Patel and her colleagues who used propositional analysis and production systems as a means of investigating differences between novices and experts. Without doubt, there are other paradigms that could be used to investigate the structure of clinical reasoning.

For example, Lemieux and Bordage (1986), Bordage and Lemieux (1991), and Lemieux and Bordage (1992) have investigated clinical reasoning from a structural semantics perspective. This approach considers the meaning contained in discourse along two dimensions. The linear (syntactic) axis refers to the way words and phrases are arranged in a sentence while the vertical (semantic) dimension corresponds to the organization of declarative knowledge into various levels of meaning and abstraction. For Lemieux and Bordage, the meaning given to the patient's signs and symptoms and to the medical grammar associated with diagnoses can be classified in terms of binary oppositions (e.g., congenital/acquired, internal/external, acute/chronic, benign/malignant). Possible diagnoses accounting for a patient's presentation are organized according to each polar term (e.g., evoking diagnoses that are congenital as opposed to those that are acquired). Using structural semantic analyses, Lemieux and Bordage (1986) investigated the reasoning of nine second-year medical students and five expert neurologists presented with a paper case of a patient with cervical arthrosis. The results indicated that successful problem solvers, which included both students and neurologists, used a greater number of distinct semantic axes. They also organized the signs and symptoms into clusters signifying the relationships between the clinical features and the semantic properties. These findings were taken as indicators of a broader and deeper understanding of the case. Subjects who were unable to resolve the problem successfully did so for a number of reasons. Some were unable to integrate the signs and symptoms because of lack of knowledge; others generated many diagnoses usually as the signs and symptoms were read. The latter behavior is similar to the ordered exclusion approach discussed by Ramsden *et al* (1989).

In most investigations of the structure of clinical reasoning, the focus is on differences between novices and experts. According to Schmidt, Norman, and Boshuizen (1990), one of the shortfalls of the research on novice-expertise differences is that no attempt has been made to synthesize the findings into a unifying model. On the basis of the results of the studies on novice-expert problem solving, Schmidt *et al* propose a stage theory of clinical reasoning.

A Stage Theory of Clinical Reasoning

In addition to considering cognitive information processing, the stage theory of reasoning, according to Schmidt *et al* (1990), considers the sequence and influence of education and experience. The theory rests on three assumptions. First, as students progress through their medical education and training, they go through several transitory stages of knowledge structure. Second, these knowledge structures remain available for use whenever the need arises, and third, experienced physicians make use of knowledge structures referred to as illness scripts. These scripts are based on the physician's practice and as such contain very little knowledge about pathophysiological causes and a whole lot about the clinical manifestations of diseases.

In the first stage, students develop elaborate causal networks or structures which explain diseases in terms of underlying pathophysiological processes. One of these structures, referred to as propositional networks, indicates how objects, events, or concepts are related. Such networks are derived from students' thinking by asking them to recall their knowledge about a particular topic, patient problem, or an event under study (for an example, see Groen & Patel, 1988). As the students acquire more education and experience, their recall protocols and, therefore, the causal networks become more and more complex. Because most of the students' learning is from textbooks and contact with patients is limited, the understanding of diseases and the way diseases manifest in patients is prototypical. Given a problem, a student's understanding and explanation of what is going on reflects a textbook description of the disease process. At this stage, the variability with which the same problem may appear in different patients has not been internalized by the student because of little or no patient exposure.

With additional education the causal networks of extensive pathophysiological understanding of diseases are refined to models which explain signs and symptoms in terms of diagnostic labels. This constitutes the second stage. These simplified networks appear as students are exposed to patients. With increasing patient contact, the networks begin to lose some of their pathophysiological framework; the declarative schema become compiled or chunked into units which consist of the knowledge pertinent to understanding the specific problem with which the patient has presented. The reasoning is contextual and only that knowledge required to resolve a patient's problem is retrieved rather than all the knowledge associated with the disease.

The third stage is characterized by a transition from causal structures to list-like structures (illness scripts) consisting of the different contextual features which characterize the clinical appearance of diseases. These scripts contain little pathophysiological knowledge about the causes of signs and symptoms but a wealth of clinical knowledge about diseases, consequences, and the contexts under which illness develops. The scripts are derived from extended practice and patient contact. The information in a script appears in a certain order which resembles the order in which physicians communicate with other physicians about patients. The script begins with enabling conditions or factors which make a certain disease more likely. This is followed by a statement of the fault -- a description of the malfunction or the problem. Faults are followed by consequences. Typically, faults are expressed as diagnostic labels. Consequences are the signs and symptoms that are associated with a fault or diagnostic label.

The final stage of clinical reasoning according to Schmidt *et al* (1990) is script instantiation. When a patient's health wanes, the physician searches his or her memory for an appropriate script to match the patient's complaints. In the course of the in-coming patient script's verification, the script is instantiated and remains in memory as traces of previously analyzed patients. Scripts of previous patients remain in memory and are retrieved in the evaluation and management of future patients. The knowledge structures acquired during the different stages of pathophysiological schema, compiled networks, illness scripts, and instance scripts do not decay but form layers in memory through a sedimentation process and each layer is available should the more refined layer fail to produce an explanation of the patient's health status.

The reasoning displayed in this final stage, according to Ridderikhoff (1991) is inductive. For the inductive thinking physician, the patient's presenting complaints trigger one or two ideas about the patient. Because of the physician's experience, a number of patients presenting with the same or similar complaints has been stored in the physician's memory. These instantiations of previous patients serve as a basis for verifying the presenting patient's problem. Verification is accomplished through questioning and, depending upon the complexity of the problem, additional information in the form of physical examination and laboratory tests may be required. The small number of generated hypotheses all serve as possible solutions. However, only one of the hypotheses is selected as most likely.

In summary, the stage model of clinical reasoning begins with the students' understanding of pathophysiological processes. As the students acquire more knowledge in the context of clinical problems, their understanding of illness shifts from elaborated networks to abridged causal ones relating signs and symptoms to diagnoses. With the introduction of patients with altered health status, illness scripts are developed for each patient condition. As more patients are seen and more illness scripts are matched, the scripts are instantiated in memory. When a patient presents with a problem, these scripts are retrieved and the physician may engage in a search and pattern-matching activity to find a script which is identical to or best approximates the patient's problem.

Successful Problem Solving and Expertise

Many of the differences between experts and novices are similar to the characteristics that differentiate successful problem solvers from unsuccessful ones. From the investigations of problem solving in fields such as medicine, biology, physics, chemistry, mathematics, programming, and electromechanical troubleshooting, Smith (1991) developed a definition of the term "problem" and identified behaviors of successful problem solvers. For Smith (1991), a problem is a task which requires analysis and reasoning. This element of a problem as well as other elements are reflected in the pictorial definition developed by Smith and which is reproduced in Figure 1.

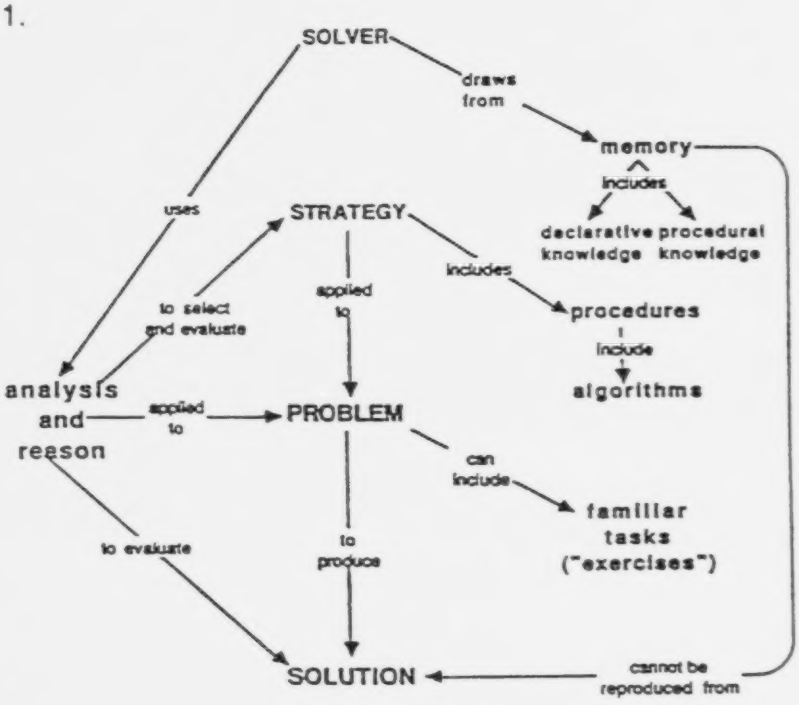
SOLVER
draws from
memory
includes
declarative knowledge
procedural knowledge
uses
STRATEGY
includes
procedures
include
algorithms
applied to
to select and evaluate
analysis and reason
applied to
PROBLEM
can include
familiar tasks ("exercises")
to evaluate
to produce
SOLUTION
cannot be reproduced from

Figure 1: Definition of the term "problem" (Smith, 1991)

To solve a problem, a problem solver has to access information from the declarative and procedural knowledge that constitutes the domain from which the problem is drawn. For Smith, a resolution of the problem can not be obtained by recall or recognition only. It would appear that in addition to recall and recognition, problem solving should involve analysis, reasoning, and the use of strategies. Successful problem solving is characterized by several features. First, successful problem solvers adapt their knowledge and its organization and apply these to facilitate a resolution of the problem. Successful problem solvers use their knowledge to create an internal problem space in which the salient features of the problem are represented or redescribed to formulate a personal understanding of the problem. The existing problem may be cast as an instantiation of previous experiences and problems. Recognition of this type helps trigger related chunks of content which along with the declarative and procedural knowledge guides the problem solving process. In solving a problem, successful problem solvers plan the general strategy or approach. Smith also states that successful problem solvers use strategies such as breaking a problem down into parts and then use either forward reasoning if the problem is in their domain of expertise or backward reasoning if the problem is not in their domain of expertise. Successful problem solvers also use other strategies such as constructing diagrams, writing out definitions, and estimating the answer. Depending upon the degree of problem complexity, successful problem solvers can perform multi-step activities and keep track of the results of each activity. Finally, successful problem solvers check and evaluate their solution for accuracy and reasonableness.

Closely related to the above characteristics of successful problem solving are the uniformities in expert performance across different domains cited by Holyoak (1991). Compared to novices, experts are more accurate in performing complex tasks. They solve problems with greater ease and have superior memory for information related to their field. Experts are better at perceiving patterns in data. Their expertise develops from knowledge which was initially acquired using backward chaining but with increasing practice, the degree of expertise increases and problems are resolved through forward rather than backward reasoning. According to Holyoak (1991) and Smith (1991), the expert-novice differences and the characteristics of successful problem solvers are predicated on the individuals' knowledge of the domain. If this knowledge is adequate, organized, and integrated then problem solving is enhanced.

Summary

In medical education, the 1970's saw the introduction of cognitive psychology to the study of medical problem solving. Using stimulated recall procedures, Elstein *et al* (1978) had 24 physicians interview three simulated patient actors. On the basis of the physician protocols, Elstein *et al* proposed that much of the physicians' reasoning in the physician-patient encounter could be explained by the hypothetico-deductive method of reasoning.

There was no doubt that cognitive psychology and, in particular, information processing were beginning to make an impact on medical problem solving research. Research in medical problem solving shifted from the investigation of general problem solving in complex, poorly designed environments to problem solving which was domain-dependent. The 1980's and early 1990's witnessed a proliferation of research on the use of knowledge in problem solving and novice-expert differences. Part of this was attributed to the finding of early hypothesis generation and the use of the deductive method of problem solving reported by Elstein *et al* (1978) and Barrows *et al* (1978). Researchers such as Patel and her colleagues began to question whether all problem solving was hypothetico-deductive and, as a result, a body of literature on novice-expert differences emerged. Much of this research focussed on the pattern of reasoning used by novices and experts. Complementing the research on expertise were additional endeavors in clarifying the clinical reasoning construct.

The collaborative research of Patel, Norman, and Schmidt (Groen & Patel, 1988; Kaufman & Patel, 1991; Norman, Tugwell, Feightner, Muzzin, & Jacoby, 1985; Norman, Brooks, & Allen, 1989; Norman, Brooks, Allen, & Rosenthal, 1989; Patel, Groen, & Frederiksen, 1986; Patel & Groen, 1986; Patel, Groen, & Scott, 1988; Patel, Groen, & Arocha, 1990; Patel, Evans, & Kaufman, 1990; Patel, Groen, & Norman, 1991; Schmidt, Hobus, Patel, & Boshuizen, 1987; Schmidt, Norman, & Boshuizen, 1990) exemplifies the work on expertise and problem solving in medicine. Others, such as Barrows and Feltovich (1987), Elstein, Shulman, and Sprafka (1990), Swanson and Stillman (1990), and Ridderikhoff (1989; 1991), have clarified the role of hypothetico-deductive reasoning in medicine.

There is evidence from medicine, as well as other fields, such as physics, that forward reasoning is used by experts and by successful problem solvers working in their own domain; experts revert to backward reasoning whenever a problem with which they are unfamiliar, or outside their domain, arises. Novices, because of their limited knowledge and experience tend to use backward reasoning.

Testing and Thinking Aloud

Researchers such as Haladyna and Downing (1989a, 1989b), Millman and Greene (1989), and Osterlind (1989) conclude that there is no comprehensive body of relevant research and literature pertaining to the construction of multiple choice test items. What does appear repetitively in textbooks are lists of rules such as "make all the alternatives about the same length", "Use 'None of the above' sparingly", and so on.

With respect to cognitive processing, Bloom's (1956) taxonomy has been used widely to label the cognitive processes required in test items. The intention of Bloom and his colleagues was to describe the different categories of mental processing in the cognitive domain with the hope that the taxonomy would impact upon instruction and assessment and that the taxonomy would help with understanding the underlying cognitive processes in responding to test items. Although many examination items were classified according to the levels of the taxonomy, such categorizing was usually cursory. Because of the constrictive nature of the taxonomy, along with the somewhat incomplete descriptions of the "higher" cognitive levels, the taxonomy fell short in providing insight to the processing of test items.

Several studies have been conducted investigating the classification of items according to categories of Bloom's taxonomy. The following section considers some of these studies.

Classifying Items

Using three subject matter experts in each of four medical disciplines, McGuire (1963) asked the content experts to classify items appearing on an examination of the National Board of Medical Examiners according to a modified version of Bloom's Taxonomy. Each expert was asked to consider each item independently and to determine, by introspection, the mental process a student would require in answering the question. Unanimous agreement was achieved by the item raters on 61% of the 683 items. Some of the items were difficult to classify because the items were constructed on the basis of content rather than cognitive processing specifications. Over half of the items were categorized as isolated recall. Because the objectives of medical education emphasize problem solving, McGuire and her colleagues at the University of Illinois College of Medicine modified their original taxonomy by combining the levels of analysis, synthesis and evaluation into a single level which they labelled as problem solving. The authors then proceeded to develop examinations and test items which were more aligned with the categories of the revised taxonomy. Along with some preliminary results from a think aloud exercise, McGuire concluded that it was possible to construct examinations which tested more complex cognitive processes.

A similar study was conducted by Cox (1978). Using a three-level taxonomy of knowledge, comprehension, and problem solving, Cox had three surgical instructors and six senior students classify 150 items. With respect to the allocation of the items to the three taxonomic levels, the instructors agreed on 44% of the items. Student agreement was lower with the students classifying more of the items belonging to the problem solving level. Cox hypothesized that the taxonomy was hierarchical and represented increasing levels of difficulty. Such was not the case; Cox had difficulty demonstrating hierarchical structure and test scores were similar across the three levels. Cox concluded that classification of test items was non-productive and that problem solving should be assessed with tests other than those consisting of multiple choice items.

Another study similar to those of McGuire and Cox was conducted by Stoker and Kropp (1964), in which the problem of whether judges were able to agree on the cognitive process that an item intended to assess was investigated. Five judges rated 36 multiple choice items on Atomic Structure according to the categories of Bloom's Taxonomy (1956). Four different judges rated another 36 items that dealt with relationships between surface area, cross-section, and weight. For the items dealing with Atomic Structure, 11 of the 36 items were classified congruent with the processing categories by all five judges. For nine items, there was agreement by four raters and for 14 items there was agreement by three raters. The findings were similar for the items dealing with surface area, cross-section, and weight. Perfect agreement was obtained for 11 items while for 16 items agreement was for three of the four raters. Stoker and Kropp concluded that the taxonomy serves as a basis for constructing items and that raters can assign items to the different categories with some accuracy.

Stoker and Kropp also investigated the hierarchical structure of the taxonomy using factor analysis and simplex structure in the correlation matrix. Some 20 different data sets were developed on the basis of different sample sizes for the two different tests. The results of the factor analysis did not support the hypothesized hierarchical structure and none of the 20 correlation matrices met the requirements of simplex structure.

In the above studies, the finding that subject matter experts were able to categorize items according to Bloom's taxonomy (or some modification of it) and that this could be accomplished with some degree of agreement does not confirm that responding to the items would in fact evoke the intended processes. As Stoker and Kropp point out, lack of agreement in classification could arise because an item writer may design an item to evoke a behavior such as "Analysis", but students might answer the item by systematically eliminating distractors. In such a case, the intended process differs from the obtained process.

Thinking Aloud and Classroom Tests

Although Stoker and Kropp made no suggestions as to how the obtained process may be investigated, McGuire (1963) did attempt to provide some confirmatory evidence for the classification tasks. To determine the degree of congruence between the cognitive process actually used by students and the predictions made by the content experts as to what cognitive process students would use in answering an item, McGuire had 30 medical students think aloud as they worked through a sample of items. Although McGuire was unable to match students' reasoning with that postulated by the subject matter experts because of interviewer inadequacies, she nevertheless found that better students often attempted to figure out an answer by reconstructing the relevant concepts. On the other hand, weaker students responded usually through visual or auditory recall of a page in a text or a section in a lecture. McGuire also stated that future studies incorporating think aloud methods should consist of a larger number of students and variation in item content, question format, student preparation, and student ability.

The use of think aloud methods to uncover cognitive processing is not new. More than 100 years ago when psychology was being established as a discipline, the dominant method of investigation was eliciting introspective verbal reports (Ericsson, 1987; Afflerbach & Johnston, 1984). Behaviorism put a damper on studying cognitive processing and a decrease in the use of verbal reports. With the advent of cognitive science (Snow & Lohman, 1989), interest in using verbal reports has been rekindled.

The McGuire study was not the first to have examinees think aloud as they answered test items. At least two studies using multiple choice items and verbal reports were reported in the early 1950's. The first of these was a study conducted by Bloom and

Broder (1950) in which a sample of college students was asked to think aloud while working on problem solving items. The following is an example of a problem solving item taken from Bloom and Broder.

> Rank the following life forms in the order of their appearance in the geologic record. Blacken the answer space A for the one that appeared first,etc.
>
> 81. Mammals
> 82. Sponges
> 83. Amphibians
> 84. Fishes
> 85. Flowering plants

A variety of items from different subject areas were used and each item was completed by six to eight students. To capture the essence of the verbal reports the authors focussed on the student's understanding of the problem, understanding of the ideas in the problem, general approach to the answer, and attitude toward the solution of problems. Students displayed differences on each of these major headings within the same item and across items. For example, general approaches to the answer revealed that successful students read the statement in the problem and generated hypotheses to the correct answer or established criteria which would be met by the correct answer. With unfamiliar vocabulary, the successful students made assumptions about the meaning of the word and then proceeded to answer the item on the basis of the assumed meaning of the unknown word. Unsuccessful examinees adopted a passive stance in solving problems; they seldom considered the task presented in the item, ignored unfamiliar vocabulary, and selected answers on the basis of impression or feeling. Answers were also selected on the basis of guessing or resorting to some ill-conceived approach such as selecting response D because on previous examinations, alternative D was the correct answer for many of the items.

In general, the verbal reports confirmed the students' solutions to the problems on the basis of the thought processes. That is, good thinking and reasoning tend to result in correct responses, whereas deficient thinking produced poorer or incorrect responses. Furthermore, a variety of methods were used by the students in arriving at the answers.

In the second study, Kropp (1956) ascertained the processes which 54 students used in answering 10 multiple choice questions selected from a standardized test that dealt with reading comprehension of passages in the physical sciences. The items were such that more than recall and recognition were required to answer the questions. Students were tested individually and their verbalizations were phonographically recorded. In matching the choice of correct answer and the verbal protocol, Kropp found that students used a variety of "processes" (Kropp's choice of word) to arrive at an answer. For some items, students translated the problem into a context with which they were familiar; for other items, answers were obtained by distractor elimination, although, the manner in which this was done differed. Some students selected the correct response and then eliminated the distractors; other students selected an answer by eliminating all but one possible response. Several students guessed correctly while others responded correctly through compensating errors such as being unable to read decimals and not being aware of inverse relationships between frequency and wave length. Others were unable to formulate appropriate computational processes but because of their sensitivity to the magnitude of numbers, they were able to respond correctly. Finally, and not a strategy, one item was given three distinct interpretations all of which led to distinct reasoning that led to the same correct answer. Because some of these processes were justified and others were not, Kropp concluded that it would be hazardous to infer the quality of the thinking processes on the basis of answers selected. He also concluded that thinking aloud while answering test questions revealed item ambiguities, hidden cues, and the variety of processes students use in answering questions which seem to have simple and straight-forward answers.

Thinking Aloud and Standardized Tests

The results from the studies by Bloom and Broder (1950), McGuire (1963), and Kropp (1956) support the supposition forwarded by Stoker and Kropp (1964) that the manner in which examinees respond to test items is not necessarily the process envisioned by the item writer. While the above studies made use of classroom examinations and test items, several studies using protocol analysis have been conducted using items from standardized examinations.

Using 13 verbal analogy and 12 antonym items taken from the College Entrance Examination Board's Scholastic Aptitude Tests, Connolly and Wantman (1964) asked nine college sophomores to report all their thoughts as they worked through the items. These think aloud reports were tape recorded. To improve rater agreement between the two raters used in the study, Connolly and Wantman developed a list of statements which described, in behavioral terms, the reasoning process. Examples of such statements included skips unfamiliar terms, justifies choice with incorrect logic, reads stem and immediately states keyed response, and appears to be guessing between two possibilities. One hundred and seventeen such statements were initially produced and, following review, the list was reduced to 67 statements. Ratings were recorded using a 7-point scale. The authors did not provide an example of a statement cast on a 7-point scale nor did they explain the scale but it would appear that the scale reflects "importance", "relevance", or "reasoning". As each rater listened to an examinee's reasoning to an item, the rater checked those statements from the list of 67 that described the student's approach. The checked statements were then used as guides in arriving at a final rating, although the rater was not bound by the scale values of the statements. Thus, for some items, an examinee may have selected the correct response and obtained a low rating while for other items a high rating may be associated with an incorrect response. A three-way analysis of variance was conducted with the two raters, nine subjects, and 25 items treated as main effects. Results of the ANOVA indicated that all three main effects and three first-order interactions were statistically significant at an alpha level set at 0.05.

Of the 67 statements describing reasoning behavior, 14 were never used, eight were used by one rater and not the second, and 10 were used by the second rater and not by the first. Fourteen statements were used frequently by both raters. In the concluding section of the article, Connolly and Wantman list a series of future activities, two of which are still as relevant today as they were in 1964. One of these dealt with the identification and classification of stylistic differences in students' approaches to test items. The authors believed that such styles existed and that verbalizing test behavior could evidence such styles. The second, which is the crux of validation, referred to relating the item designers' intentions of the measurement properties of the items to what actually happens when examinees respond to the items. This information could then be passed on to other item writers.

A more recent study using verbal reports and items from standardized tests was reported by Haney and Scott (1987). Concerned about ambiguous items appearing on standardized tests, Haney and Scott conducted a pilot study using 29 items in social studies and science and 32 items in reading sampled from standardized examinations such as the Metropolitan Achievement Test, Comprehensive Test of Basic Skills, California Achievement Test, and the Stanford Achievement Test. Eleven second- and third-grade children participated in the study. Following group administration of the items, each child was interviewed individually about each test item. Interviews were audio-recorded. In the interview for the social studies and science items, each child was asked to describe the alternative answers and then, after the interviewer repeated the test question, to explain the reason for selecting a particular response. For the reading items, each examinee was asked to read the question and the alternatives aloud and then to explain reasons for answer selection. The audio-recorded interviews were transcribed and a rating scheme was developed which dealt with the aspects of answer selected, perception of item stimulus, perception of response alternatives, process of reaching an answer, reasoning, and extraneous knowledge.

The results indicated that only in one instance was a child rated as to having perceived an item in a way unrelated to the skill measured by the item. Of particular interest were the findings related to process, reasoning, and extraneous knowledge. For the majority of the items (around 90% for each of social studies, science, and reading), students evidenced some form of reasoning or logic. The second most common strategy of arriving at a correct answer was guessing, followed by elimination of alternative answers and recognizing a mistake which was recorded on the examination but was self-corrected during the interview. Valid reasoning was displayed in 80% of the reading comprehension items, 73% of those in science, and 68% of the social studies items; extraneous information was only used in about 2% of the items. Item ambiguity, defined as instances in which a keyed response was selected for reasons unrelated to the skill measured by the item, was apparent in about one-third of the items on the three sets of items. Marking an unkeyed response, even though the reasoning was a valid application of the skill, appeared in 8% of the reading comprehension items, 3% of those in science, and 7% of those in social studies. Haney and Scott conclude that there is a need for input from examinees as part of the process for developing and validating test items.

Identifying Strategies

Thus far, the review of the literature has dealt with agreement in item classification according to taxonomic levels and with the use of think aloud methods to identify cognitive processes and strategies students employ in responding to items on teacher-made and standardized tests. More concerted efforts in identifying strategies have been carried out with young children in the subject area of arithmetic. In the study of strategies used by 4- and 5-year old children in the addition problems of the type "If you had *n* oranges and I gave you *m* more, how many would you have then?", Siegler and colleagues (Siegler & Robinson, 1982; Siegler & Shrager, 1984; Siegler, 1986, 1989) identified four strategies. In these simple problems where *n* and *m* could range from 1 to 5, children sometimes raised their fingers and counted them (the counting fingers strategy) and other times they raised their fingers and answered without counting (fingers strategy). Simply counting aloud without an external referent such as fingers was another strategy and finally some children answered without any intervening overt behavior between presentation of the addition problem and the response. The last strategy was labelled as retrieval. Interestingly enough, the same four strategies were used in subtraction problems (Siegler, 1986).

To account for the different strategies used by the pre-school children in solving addition and subtraction problems, Siegler and Shrager (1984) developed the distributions of associations model. The model consists of two parts, a representation of knowledge about the problems and a second part that deals with the process that operates on the knowledge representation to produce answers. Children associate correct and incorrect responses with specific problems. For example, the addition problem, 3 + 4 =, would be associated with the correct response 7 as well as the incorrect responses of 5, 6, and 8. Each of the responses, 5, 6, 7, and 8 would have a numerical value referred to as an associative strength. These associative strengths are analogous to the proportions of examinees selecting each alternative as a possible answer in a multiple choice item and would be derived from previous test administrations. The process part involves three sequential phases, any one of which could produce an answer to a problem. In the first phase, an answer is retrieved. If the answer's associative strength exceeds the examinee's confidence criterion, then the retrieved answer is stated. If the examinee is not sufficiently confident of the retrieved answer, the next sequential phase is one of elaboration in which the examinee might represent the problem by putting up fingers and trying to recognize the number of fingers. If the examinee still lacks confidence in the answer, the examinee will apply an algorithm (final phase) such as counting the raised fingers and stating as the answer the number corresponding to the last count.

In the multiplication of two numbers where the multiplicand and multiplier could range from 0 to 9, Siegler (1986) found that children used both overt and retrieval strategies. Two types of overt behavior were identified. One was an elaboration of the problem where the children would either repeat the multiplication task orally or write the problem down on paper. The second overt behavior consisted of elaboration of the problem and computation. For some problems, the multiplicand would be written the number of times indicated by the multiplier and then added. For other problems, a literal representation and counting was used. Thus, 4 X 2= __ was represented as 4 bundles each having 2 lines and the answer was obtained by counting the number of lines.

The strategy of diagraming a multiplication problem and thinking aloud has implications for medical examinations consisting of multiple choice items which incorporate pictorial material. For example, a photograph of a skin lesion may be accompanied by an item requesting choice of treatment. An underlying assumption in such an item is that to prescribe treatment requires a correct diagnosis. Thus, if an incorrect response is provided by an examinee, it is not known whether the examinee experienced difficulty with only treatment or with diagnosis or both. Having examinees think aloud for such items would clarify the sources of difficulty.

Siegler's other work on strategy choices in spelling likewise shows that children use a variety of behaviors to spell a word. These include looking for the word in a dictionary, sounding the word out, retrieval, and writing out alternative spellings and selecting one as being correct.

In a more recent study dealing with the identification of strategies, Farr, Pritchard, and Smitten (1990) examined the criticism that on multiple choice tests of reading comprehension, examinees do not read and comprehend the reading passages that accompany the test items. Twenty-six college students were assigned randomly to either an introspective or to a retrospective group and asked to respond to 24 multiple choice questions related to 3 reading passages taken from the Iowa Silent Reading Test. The interviews were tape-recorded, transcribed, and summarized. Each summary was analyzed for overall strategy, shifts in overall strategy, specific reading and test-taking strategies, and difficulties reported by examinee.

Results indicated that retrospective and introspective groups performed similarly (72.4% and 72.6% respectively) on the 24 items. Four overall strategies were identified by Farr *et al.* These were 1) the passage is read and then each question is read and is followed by a search of the passage for the correct answer, 2) the passage is partially read and then each item is read and is followed by a search of the passage for the correct response, 3) all items are read and then the entire passage is read followed by a rereading of each item and a search of the passage for the correct answer, and 4) the first item is read and is followed by a search of the passage for the correct response and the pattern is repeated for each item. Of these, the first overall strategy was used by 62% of the 26 examinees. In completing the three passages seven examinees shifted from one strategy to another. Specific strategies noted while examinees read the passages included determining the theme or purpose of the passage, making predictions, noting location of information either mentally or by marking the text, predicting the questions which may be asked, and using test questions to determine theme or purpose. The strategies used by the examinees in answering the test items were grouped into meaning centered and non-meaning centered. Meaning centered strategies included rereading items to clarify, generating an answer before looking at the alternatives, thoughtful consideration of alternatives, and looking back at the passage for specific words or phrases. Non-meaning centered strategies included eliminating an alternative when the exact word was not found in the passage and not choosing an alternative either because of its location or because of unfamiliarity with its meaning. In reading the passages and the associated questions, examinees used metacognitive strategies such as adjusting pacing, holding off on an answer, adjusting overall strategy, and increasing concentration. Farr *et al* concluded that the reading passages and test questions were intertwined and that the strategies used by the examinees is an estimate of reading ability and therefore supports the construct validity of multiple choice reading comprehension tests.

In a study conducted by Norris (1990) and which will be described in more detail in the latter parts of Chapter II, Norris performed a qualitative analysis of the verbal reports of 40 randomly selected examinees who were responding to items on the Test on Appraising Observations. From the analysis of the verbal reports, seven reasoning strategies were identified. These strategies included citing factual details, self-questioning, making evaluations, constructing supporting assumptions, controlling attention, interacting with the experimenter, and pausing.

Quality of Thinking

The use of think aloud methods to gather information is increasing and continues to be used in areas such as spatial ability, reading, mathematics, problem solving, and critical thinking. The emphases in these studies appears to be on identifying strategies and determining the relationship between selection of keyed responses and quality of reasoning. In addition, research which addresses some of the issues related to methodology of verbal reports is being conducted, although such studies are few.

One of the early studies that examined quality of thinking was done by Schuman (1966). He had 1000 factory workers and cultivators in East Pakistan complete an attitude survey and then report on the reasons for their responses to 10 randomly selected items. The subjects' reasons were rated on a scale of 1 (reason is clear and leads to accurate prediction of correct response) to 5 (reason is very unclear and could not be used to predict the correct response). Mean scores were calculated for each item and used to indicate the subjects' understanding of that item. Schuman's study is a departure from the earlier studies because it represents the first attempt to quantify verbal protocols.

More recently, Phillips (1989) developed a Test of Inference Ability (TIA) in reading comprehension. The TIA consisted of three reading passages, each accompanied by 12 scaled answer multiple choice questions. That is, the four alternatives account for different levels of understanding and inference and numerically take on values ranging from 0 to 3. An examinee response that is consistent with the information presented in the text and, likewise, consistent with the examinee's prior or background knowledge is worth three marks; a partially correct response is worth two; a response based on the text alone is worth one mark and an incorrect response is assigned a score of zero.

Part of the validation of the TIA included the use of student verbal reports to determine whether a complete inference had been made and whether there were any item ambiguities, vocabulary problems, and hidden cues. Verbal reports were also used as a basis to determine the relationship between answer selections and thinking process. Verbal reports clarified several aspects. First, the range of responses was extended for some items on the basis of degrees of inference displayed in the think aloud protocols; second, words presenting problems in understanding were replaced by words and/or phrases, and third, problematic areas such as time frame shifts which presented students with difficulty despite repeated revisions were either replaced or deleted from the examination. For the majority of the items comprising the TIA, good thinking correlated with good inference making while poor thinking was related to poor inference making.

Along lines similar to that of the Phillips' study, Norris (1989a, 1989b) investigated the use of verbal reports on a multiple choice test of critical thinking referred to as the Test on Appraising Observations. An example of an item taken from the test appears below.

> A policewoman has been asking Mr. Wang and Ms. Vernon questions. She asks Mr. Wang, who was one of the people involved in the accident, whether he has used his signal.
> Mr. Wang answers, "Yes, *I did use my signal.*"
> Ms. Vernon had been driving a car which was not involved in the accident. She tells the officer, "*Mr. Wang did not use his signal.* But this didn't cause the accident."
> (Norris, 1989a, p. 6)

The task for the examinee is to indicate which, if either, of the italicized statements is more credible. According to Norris, the difficulty presented in scoring such an item is no different than the scoring of, say an item on an achievement test. It is possible to answer an item correctly through faulty reasoning. In a multiple choice examination, a way around this problem is to request for verbal reports. For the Test on Appraising Observations, Norris attempted to ground the construct of judgment by providing an ideal model of thinking. For the automobile accident item, ideal thinking would reflect the following components.

> Mr. Wang was involved in the accident, but Ms. Vernon was not involved. Mr. Wang is less credible because his involvement would give him reason to say he used his signal even if he did not. Wang is in a conflict of interest. People in a conflict of interest, that is, people who have something to gain by events being cast as they described them, tend to be less credible than those who are not in such a situation (Norris, 1989a, pp. 6-7).

The concern in tests of judgment and critical thinking is that the keyed answer represents the judgment of the test designer. This judgment is usually based on a number of considerations, criteria, ideologies, and beliefs that a test designer brings into the test and which hopefully match the ideologies and beliefs that examinees bring into the test situation. The problem is that the two are sometimes at odds and there is no evidence as to what examinees consider when responding to judgment and critical thinking examination items. For each item, an examinee was assigned a performance score according to whether the examinee agreed (1) or disagreed (0) with the keyed response. In addition, each examinee was assigned a thinking score (ranging from 0 to 3 but later reduced to 0 to 2) derived from the verbal report. The weighting for the thinking scores were based on model responses. A thinking/performance index was then developed for each item as follows. Combinations (0,0) and (1,1) both yield a weight or a thinking/performance index of 1 because in the first case poor thinking is associated with selecting a wrong response; in the second case good thinking is associated with a correct response. The two remaining combinations both yield a negative weight because (0,1) represents the wrong answer but good thinking whereas (1,0) represents the correct answer but poor thinking. Using this index and the biserial coefficient, Norris made revisions to the items.

Verbalizing and Thinking

With respect to issues of methodology, Garner (1988), Ericsson and Simon (1980), and Nisbett and Wilson (1977) have argued that think aloud procedures could disrupt the metacognitive process to generate the verbal report and alter the elicited strategic information. A necessary condition for think aloud protocols to be relevant to the validation task is that the verbal reporting not alter performance and thinking. To investigate this necessary condition, Norris (1990) used the Test on Appraising Observations. A total sample of 342 students in grades 10, 11, and 12 were assigned randomly to one of four elicitation groups or to a control group. The four elicitation groups consisted of 1) think aloud, where subjects were instructed to report all they were thinking as they worked through test items, 2) immediate recall, where subjects were instructed to record their answers to each item and then tell why they selected the answers, 3) criteria probe, where the examinees were asked to record their answers and then were asked whether a piece of information pointed out in the item had made any difference on their choice of response, and 4) principle probe, where subjects were treated like the criteria probe group with an additional question asking whether their answer choice was based upon particular general principles. The control group received no elicitation; they were simply instructed to work alone and to record their answers on the answer sheet. Verbal protocols were tape-recorded for items 1-15 for the four elicitation groups. The remaining 13 items were completed in a paper and pencil format with no verbalizations. Three scores were generated for the four elicitation groups - a performance score on the first 15 items, a score on the remaining 13 items, and a thinking score. Identical scores except for the thinking score were generated for the control group.

The mean scores for the four elicitation groups and the control group on the first 15 items ranged from 7.6 items correct for the principle probe group to 8.3 items correct for the immediate recall group. The think aloud group had a mean of 8.0 items correct while the control group answered on average 7.8 items correctly. On the subsequent 13 items, means ranged from 8.1 to 8.6 items correct. Thinking scores for the four elicitation groups ranged from 7.9 (think aloud) to 9.2 (immediate recall). Norris concluded that verbalizing to the examination items did not alter the examinees' performance and thinking scores. The verbal reports of thinking on the multiple choice examinations did not change examinee thinking and performance from what it would have been had the examinee taken the examination in a strictly paper and pencil format. Probing for reasons or requesting examinees to justify their choices altered neither their thinking nor their performance. Likewise thinking and performance was not affected by asking examinees directed and leading questions which focused on specifics. The finding that the act of thinking aloud does not alter the thinking and performance lends some credibility to the use of verbal reports as a way of providing additional evidence for validity of test items.

The effect of thinking aloud on reasoning is not the only concern; several other methodological issues related to the implementation of think aloud procedures will be discussed in the next chapter.

Summary

To summarize, at the core of any examination is the test item. Responses to test items are interpreted as indicators of some underlying cognitive process. Inferences about these cognitive processes are usually inferred because the measures are indirect. Providing item writers with better item specifications and taxonomies has helped increase the confidence in the inferences. However, such inferences will always possess a degree of suspicion unless there is some evidence which shows that the evoked response matches the item writer's intent. Research on these matters has shown that providing item writers with detailed item specifications is not universal, that agreement between subject matter experts to categorize items into taxonomic levels is moderate, and that examinee's verbal reports of test-taking behavior do not always mirror the intended cognitive processes. A variety of strategies, some relevant and others not relevant to the construct of interest, may be used to arrive at a response. Thus, Norris (1989a) in discussing the use of multiple choice items designed to specifically test critical thinking concludes that construct validation should provide an explanation of performance on a test by modelling the cognitive or mental processes of critical thinking. However, in order to provide such evidence it is necessary to implement methods, such as verbal reports, to provide a more direct manifestation of critical thinking.

Because the number of studies in the field of medicine using concurrent think aloud methods of students responding to multiple choice items is small, Norris' suggestion applies equally well to tests other than those which assess critical thinking. Using think aloud methods with multiple choice items designed for assessing achievement in medicine would provide direct evidence of the process of thinking leading to the selection of responses.

The next chapter deals with the method for the present study. In addition to the method, Chapter III presents the results obtained from the initial analysis of the think aloud protocols. These preliminary results list the molecular activities or moves identified from the students' interactions with the items. In Chapter IV, the moves are combined into strategies and the strategies are examined in terms of the strategies used by experts and successful problem solvers.

III. METHOD AND INITIAL ANALYSIS

This chapter deals with the topics of selection of examinees and test items, procedure, and initial and further analyses.

Selection of examinees and items for the conduct of a study dealing with test-taking strategies within the the Faculty of Medicine at the University of Alberta presents many options. For example, examinees could be selected from either Phase I (first year), Phase II (second year), Phase III (third and fourth year), or some mix from all three phases. If first year students are used, then the test items could relate specifically to one of the ten different courses offered in Phase I or the items could be representative of all ten courses. A similar scenario arises for Phase II where the number of courses is fifteen. Phase III represents a change in a medical trainee's education; the number of courses is reduced and there is a switch from the basic sciences (Phases I and II) to the clinical sciences with an emphasis on patient care.

Selection of students and items was guided by the literature on medical problem solving. For the majority of the studies, the subjects were practicing physicians and/or students in the clinical years of training and the tasks were likewise clinical in nature. Using the literature as a guideline, three criteria were set for the selection of students and items. First, since Phases I and II are heavily laden with courses which do not have a strong clinical flavor, students from Phase III would be used. Second, items that related to a single discipline and preferably a discipline which had ward work, that is a rotation, would be used. The reason for selecting a discipline with a rotation is that the nature of bedside teaching in a rotation is not too different from the task of asking students to think aloud as they work through a set of test items. That is, the Phase III student is asked to do routine medical work-ups of patients and then to present the findings to a senior staff member. There is considerable interaction between student and staff member during the presentation of the sort, "Why did you ask about smoking habits?, "What were you thinking off when you examined the patient's lymph glands?" The interaction and the questioning is akin to the think aloud procedure. Finally, a discipline that had formal student evaluation at the conclusion of the rotation which used multiple choice items, and which had a bank of such items would be used. This would allow for a selection of items with reasonable technical characteristics.

Several disciplines met these criteria and of these, medicine was selected.

Selection of Examinees

Since the study dealt with test items in the medical field and uses students enrolled in third or fourth year in the Faculty of Medicine, approval for the conduct of the study was obtained from the Associate Dean for Undergraduate Affairs. Ensuring student participation was obtained through the president of the Medical Students Association, who was also briefed on the purpose of the study. Students were recruited for participation in one of two ways.

Students complete their rotations at either the University of Alberta Hospitals or one of the affiliated hospitals. During a rotation between four and eight students are assigned to a staff member located at one of the hospitals. Initial contact was made with the staff member requesting participation of the students and permission to meet with the students for the purposes of explaining the study. Once permission was obtained from the staff member, the investigator met with each group of students during which time the purpose of the study, participant tasks, procedures, and approximate time required to complete the task were discussed. Students were informed that participation was voluntary and if they were interested in serving as subjects to give their names and telephone numbers. Students were told they would be contacted by telephone within several days of the meeting to determine participation and appointments. In a telephone follow-up, of the 48 students who would be completing their rotations during the time period February 1 to May 29, 1991, 33 students volunteered as participants. Of the 33 students, 10 were female and 23 were male. A comparison of performance on a set of course examinations administered during the first half of Phase III showed that the 33 students were similar to the total class of 115 students. The mean performance for the 33 students was 72.0% with a standard deviation of 5.0, while the performance of the total group was 71.8% with a standard deviation of 4.3%.

In the second approach for student participation, the investigator met with third- and fourth- year students during a whole class lecture. From the eighteen students who volunteered, seven fourth year students (2 females and 5 males) accepted on telephone follow-up. The mean performance of these seven students on the Comprehensive Examination was 66.9% (standard deviation of 6.6%) while for the total class the mean was 64.4% (standard deviation of 6.7%).

As a result of the two approaches, twelve females and twenty-eight males participated in the study.

Test Items

For the purposes of the study thirty single best answer items from an existing medicine test item library were selected. All thirty items had been previously administered to a group of 118 examinees; Table 1 presents the item difficulties and the associative strengths of the distractors. The associative strengths of the distractors are simply the percent of examinees selecting each distractor as a possible answer. The correct or best answer is indicated by an asterisk beside each item's difficulty.

For item 1, which requests an examinee to select a diagnosis given some patient data, alternative 1 was selected by 1% of the examinees, alternatives 2 and 4 were not selected, alternative 3, the keyed answer, was selected by 94% of the examinee group, and alternative 5 was selected by 5% of the examinees.

Table 1

Associative Strengths of Alternatives

		Alternative				
		1	2	3	4	5
Item	Task					
1	Diag.	1	0	94*	0	5
2	Comp.	34*	12	6	2	46
3	Comp.	5	0	28	4	63*
4	Know.	77*	14	2	4	3
5	Diag.	0	84*	0	0	16
6	Test	0	23	72*	5	0
7	Test	4	82*	6	7	0
8	Know.	4	0	71*	3	22
9	Know.	4	14	24	18	41*
10	Know.	1	6	9	76*	8
11	Know.	21	72*	4	2	1
12	Comp.	75*	14	7	0	4
13	Test	1	3	76*	5	15
14	Diag.	2	39	0	56*	3
15	Comp.	2	68*	25	1	4
16	Manage	0	93*	2	1	4
17	Test	11	8	35	27	19*
18	Manage	7	42	7	33*	11
19	Manage	78*	10	10	2	0
20	Test	0	0	11	0	89*
21	Diag.	1	4	3	3	90*
22	Comp.	0	7	60*	1	32
23	Manage	91*	2	4	1	3
24	Know.	5	8	3	84*	0
25	Test	49*	9	19	17	7
26	Manage	16	79*	1	2	2
27	Diag.	3	8	1	74*	15
28	Manage	2	20	32*	46	0
29	Know.	4	4	1	28	62*
30	Manage	0	0	8	82*	10

In addition to difficulty, items were also selected using parameters such as system and task. Table 2 shows the distribution of items according to system and task.

Table 2

Distribution of Items by System and Task

	TASK					
	Know.	Comp.	Diag.	Test	Manage	TOT.
SYSTEM						
Cardiovascular	2	2	1	1	-	6
Cutaneous	1	-	-	1	1	3
GI	-	1	1	1	2	5
Immune	1	1	-	1	1	4
Musculo	2	-	1	-	-	3
Neurological	1	-	1	-	1	3
Pulmonary	-	1	1	2	2	6
TOTAL	7	5	5	6	7	30

The number of systems spanning the discipline of medicine is about twenty and the most common ones of Cardiovascular, Cutaneous, Gastrointestinal (GI), Immunology (Immune), Musculoskeletal (Musculo), Neurological, and Pulmonary were selected. Cardiology and Pulmonary were represented by six items each, GI by five, Immune by four, and the remaining three systems by three items each.

Although items in the test item library are not classified according to taxonomic levels, items requiring different types of tasks were chosen. Items were classified according to type of task by the author and chair of the Undergraduate Medical Education

Committee for the Department of Medicine. Items classified as *Know.* and *Comp.* are like Bloom's categories of Knowledge and Comprehension. *Diag.* items require students to select a diagnosis while *Test* items deal with the ordering of laboratory tests and investigative procedures. The last task, *Manage*, consists of those items in which the required activity is to select appropriate treatment and management for the patients. The distribution of items across tasks ranges from five items for *Comp.* and *Diag.* to seven items for *Test* and *Manage*.

With respect to difficulty, Tables 3 presents the distribution of item difficulty across systems while Table 4 shows the distribution of item difficulty across tasks.

Table 3

Distribution of Items by System and Difficulty

	DIFFICULTY			
SYSTEM	Difficult	Medium	Easy	TOTAL
Cardiovascular	2	2	2	6
Cutaneous	2	1	-	3
GI	-	3	2	5
Immune	1	3	-	4
Musculo	-	2	1	3
Neurological	-	1	2	3
Pulmonary	1	3	2	6
TOTAL	6	15	9	30

Item difficulty categories were assigned arbitrarily -- difficult items were those answered correctly by less than 50% of the examinee group, medium difficulty ranged from 50% to 79%, and easy items had difficulty indices which were 80% or higher. Of the thirty items, six were labelled as difficult, fifteen as medium, and nine as easy. Each system was represented by at least two categories of item difficulty. Items of medium difficulty span all systems. All tasks except for Comprehension and Diagnosis are represented by the three categories of item difficulty. Items of medium difficulty span all task categories.

Table 4

Distribution of Items by Task and Difficulty

	TASK					
DIFFICULTY	Know.	Comp.	Diag.	Test	Manage	TOT
Difficult	1	1	-	2	2	6
Medium	5	4	2	2	2	15
Easy	1	-	3	2	3	9
TOTAL	7	5	5	6	7	30

To a certain extent, the number of examinees and items used in the present study were guided by previous research. According to Afflerbach and Johnston (1984), the number of subjects in verbal report studies is relatively small. Norris (1990) randomly selected forty examinees from 342, Farr et al (1990) had twenty-six students, and most studies reviewed in Chapter II likewise used a small number of subjects and a small number of items. However, studies using medical scenarios and medical students and physicians were the exception using either one or two scenarios and generally fewer than twenty subjects.

Procedure

This section deals with the topics of tasks, training, and probing.

In order to identify and describe the strategies used by students, a think aloud protocol was used. The entire examinee-item-interviewer interaction was tape-recorded. According to Norris (1989a), it is necessary that the interviewing session be as non-leading as is possible. To allow for this, the interviewer followed the guidelines presented in the interviewing model which appears in Appendix I. In Step 1, the interviewer informs the student of the general purpose of the interview, the student's role, and that he or she will be asked to read some multiple choice items, and to respond to them verbally. Step 2 seeks information on two distinct but related activities - the manner, style, order, or approach in reading the multiple choice item and the thinking involved in arriving at a correct answer. To ensure the completeness of the think aloud protocol, it is sometimes necessary to interrupt an examinee's thinking, to seek additional information because of ambiguity, or to respond to examinee inquiries. Examples in the guidelines for conducting the interview illustrate how these situations were dealt with. Step 3 of the interview begins when the examinee has chosen the correct answer and finished reporting on the approach to the item and the thinking related to the item. The process now reverts to Step 2.

With respect to tasks, examinees are required to perform at least two tasks in think aloud exercises - responding to some activity (primary task) and verbalizing the behaviors associated with the activity (secondary task). Concern has been raised by Ericsson (1987), Afflerbach and Johnston (1984), and Ericsson and Simon (1980) that the nature of the tasks can affect the overall strategies used by the examinees. The primary task in studies using verbal reports varies from reading passages and then providing a summary or precis of the passages to filling in cloze sentences and answering multiple choice items. Preparing a precis for a passage may require more data to be held in short term memory thereby interfering with the verbal reporting or think aloud process. In such instances, the verbal reports may not be complete since some of the thinking behaviors would be lost. In the present study the primary task for examinees was to respond to a set of multiple choice items. Because responding to multiple choice items provides natural breaks between the stem and the alternatives and between the alternatives, it was assumed that the interference for examinees between responding to multiple choice items and telling what they are thinking would be minimal.

Verbal reporting may be concurrent or retrospective. Concurrent verbal reports are given while the primary task is being performed. Retrospective reports are given by the examinees after performing the primary task. An important finding of the Norris

(1990) and the Farr *et al* (1990) studies was that there were no differences in performance between concurrent and retrospective verbal reporting examinees. Thus for participants who have difficulty verbalizing during the time of the primary task, an alternative would be for such participants to reconstruct their thinking immediately following their response to the item. In the present study, verbal reporting was concurrent.

For most people, the task of verbal reporting would be new. However, as stated earlier, for third and fourth year medical students who have been on the ward, the task would not necessarily be novel since bedside teaching and seminars incorporate aspects of the think aloud method. To familiarize examinees with the tasks involved in verbal reporting, researchers have used different training procedures. The most common of these are demonstration tapes, practice think aloud sessions, and pre-study introspection (Afflerbach & Johnston, 1984). The concern that arises here is that the training procedures may bias the protocols. In the pilot study for the present study, students were read several sample verbal protocols as examples of how others interacted with the test item. A practice session where each examinee did a "think aloud" for two items was also used. Student responses to a questionnaire administered in the pilot study indicated that the sample verbal reports and the practice session did not strongly affect their approach and thinking. In the present study, both training methods were used for each student.

As mentioned earlier forty students volunteered to participate in the study. Student interviews were conducted individually over a two-month period. At the start of the interview, each student was briefed on purpose, tasks, and procedures. Appendix II provides the introduction that was used to brief the students with respect to purpose, tasks, and roles. Although all participants were familiar with the multiple choice item format in which one alternative is considered as the best answer, examples of items were provided. In addition, four different think aloud protocols for one multiple choice item were used to familiarize the students with the content and the process of thinking aloud. A copy of the four think aloud protocols was made available to each student so that the student could follow the reading. After each example was read to the student by the investigator, the protocol was reviewed emphasizing the aspects of order and parts of the question being read, thinking, selection of an answer, and justification for the the selected answer. Following this each student was given a practice item. The student's think aloud protocol was reviewed after which the student had the option of doing another practice item or proceeding to the items in the study. Most students did not seek a second practice item and instead requested to proceed with the study items.

Sample verbal protocols used to familiarize the students with thinking aloud are presented in Appendix III while the practice and experimental items appear in Appendix IV.

During the course of the pilot study, it was observed that at times students stopped reporting. In such instances students were probed with the question "Could you tell me what you are reading and thinking now?". Afflerbach and Johnston (1984) state that although probes can serve the purpose of eliciting more complete verbal reports, they can also be disruptive, interfere with the cognitive processing, and bias the reports. It was also observed during the pilot study that reasons for the elimination or selection of alternatives were not always provided. To specifically ask for a reason would constitute a probe which might alter the thinking. However, Norris (1990) has shown that requests for reasons did not alter thinking and that probing by asking students whether a specific piece of information in the item had made any difference in response selection, or whether their response selections were based on particular general principles did not affect their thinking. Based on this, it appeared that asking a specific question such as "Why do you consider alternative three wrong?" or "What are your reasons for eliminating alternative three?" would provide a more complete verbal report and at the same time not affect the cognitive processing. Thus, where there appeared to be a delay in the student's verbal reporting, students was probed with questions such as "Can you tell me where you are?" and "Can you tell me what you are thinking?". Apart from the individual probing, the conduct of the interview followed the guidelines presented in the interviewing model which appears in Appendix I.

At the conclusion of the interview, each student completed a questionnaire (Appendix V) consisting of twenty-three items that requested his/her opinion on 1) purpose of study, 2) instructions, 3) think aloud examples, 4) when thinking about the test items, 5) the exam, 6) exam content, and 7) the test items. Each of these aspects was represented by sets of adjective pairs (for example, good-bad) and opinions were specified using a four-point scale (for example, good _X_ : ___ : ___ : ___ bad). The number of students selecting each point of the adjective pair was counted and expressed as a percent.

Students were provided with the keyed answers to the items and the interview usually ended with a conversation that dealt with the role of testing and teaching in an undergraduate medical curriculum. Throughout the entire interview, ample opportunity for questions and clarification of tasks was provided. The amount of time to complete the thirty items by the forty students varied from about fifty minutes to two hours with a mean time of about eighty minutes. The total time for the entire interview averaged about two hours. On average, about forty minutes was devoted to discussing the purpose of the study, doing practice items, and debriefing.

Initial Analysis

To reiterate, this study served two purposes. The first of these was to identify and describe the strategies medical students use when responding to multiple choice items. The second purpose was to determine whether these strategies bear resemblance to the characteristics of clinical reasoning and successful problem solving.

According to Afflerbach and Johnston (1984), research into verbal reporting is a "boot-strap operation" (p. 315). That is, to determine the strategies, reasoning, and any similarities and differences in test-taking behavior across items and/or across examinees requires that the experiment run its course. The obtained information is used to identify strategies, to classify these strategies, and to identify different reasoning styles, all of which could be tested against data collected from later studies. Afflerbach and Johnston (1984) advise against forcing verbal responses and reasoning styles into predefined categories because what might appear as a disparate strategy initially may in fact turn out to be a common strategy.

With respect to the identification and description of strategies, two phases comprised the initial analysis of the student-item interactions. The first phase consisted of 1) noting global activities during the examinee-item interaction, 2) developing verbatim transcripts of the examinee-item think aloud protocols, 3) developing a list of examinee-initiated activities or moves used in responding to the items, and 4) classifying the examinee-item interaction using the list of examinee-initiated activities. In the second phase, the frequency with which each move was used was tabulated for each item across the forty students.

The task of identifying, categorizing, and tabulating the activities which emerged from the student-item interactions was based on the suggestions made by Afflerbach and Johnston (1984), Ericsson and Smith (1991), and Olson and Biolsi (1991) and reflects the procedures used by researchers such as Garner (1982), Haney and Scott (1987), Farr, Pritchard, and Smitten (1990), and Norris (1990). The next section elaborates on each of the two phases of the initial analysis.

During each of the examinee-item interactions, the interviewer noted some of the global activities occurring. Taking notes served the purpose of the interviewer being actively involved during each examinee's processing of each item. An example for the following item (ITEM ID 1034) will clarify this aspect. The best answer is indicated by an asterisk.

> The most definitive diagnostic test for pulmonary embolism with or without infarction is:
>
> 1. Perfusion lung scan.
> 2. Ventilation-perfusion lung scan.
> 3. A decreased arterial pCO_2.
> 4. An increased alveolar-arterial oxygen difference.
>
> * 5. Pulmonary angiogram.

The notes accompanying this item for one participant (examinee 01) are as follows. "Reads stem, generates 5 right off." Opposite each alternative beginning with the first one and continuing to the fifth are comments such as -- "good", "good but other things", "not definitive", "not definitive", and "definitive" respectively. The note-taking concludes with the answer selected by the examinee. Thoroughness of notes varied across items and examinees; for some items only the examinee's choice of answer was recorded.

Following the completion of the interviews, each audio-tape was transcribed. The transcript for the item presented above for an examinee appears below.

> Most definitive diagnostic test for pulmonary embolism with or without infarc, just from straight what I know the definitive is a pulmonary angiogram, the others are lung perfusion, V-Q scan and something else. So I'll look at the choices now. pCO_2 is not very definitive at all, alveolar arterial again it helps but its not very definitive, pulmonary angiogram is definitive. V-Q scan is good, it helps but it can be caused by other things as

> well, perfusion lung scan again is in the same league. Pulmonary angiogram definitive - five.

After all the audio-tape protocols were transcribed, the transcripts were grouped by item and read for an overall impression of how the examinees approached each item. In conjunction with the notes recorded for the items, the verbatim transcripts, and the findings reported for successful problem solving, a list of examinee-item interaction activities or moves was developed.

The moves were grouped into three categories. The first category (A) reflects global item-related activities. Behaviors associated with the disposition of the alternatives form the second category (B) while the third group (Category C) consists of activities associated with successful problem solving. The categories are not mutually exclusive and, therefore, some of the behaviors may be common to two or to all three categories. Each of these categories is described below.

A. Item-Related Activities

Five global item-related activities, labelled as A1, A2, A3, A4, and A5 were identified. These activities reflect the examinees' general approach to a multiple choice item. To illustrate the item-related activities, examples of students' responses to the following item will be used.

> One week after an anterior myocardial infarction a 55-year-old man complains of severe pain in the left leg. The leg is cool, pale and pulseless.
> The most likely diagnosis is:
> 1. Deep venous thrombosis.
> 2. Ruptured left iliac aneurysm.
> * 3. Arterial embolism.
> 4. A-V fistula.
> 5. Arterial thrombosis.

A1. *Reads stem and all alternatives.* This move is characterized by several activities. Examinees may read the entire item as a unit, or they may read the stem only and then provide an interpretation of what was read. They then go on to read the alternatives as a unit. The examinees may read the alternatives once or the examinees may reread all or several of the alternatives. Consider the following example of a student's first encounter with this item.

> One week after an anterior myocardial infarction a 55 year old man complains of severe pain in the left leg. The leg is cool, pale, and pulseless. The most likely diagnosis is 1. DVT, 2. Ruptured left iliac aneurysm, 3. Arterial embolism, 4. AV fistula, 5. Arterial thrombosis.

The student has merely read the item. Thus far there is no indication of what the student is thinking. Following this initial reading of the entire item, the examinee went on to provide the following interpretation:

> OK, we have a 55 year old man with acute anterior MI, severe leg pain in the left leg and the leg is cool, pale, and pulseless. The fact that the leg is cool, pale, and pulseless leads me to think about arterial problems, arterial insufficiency. His recent MI you start to think about atherosclerotic disease in the gentleman. So I'd like to think about some sort of thrombotic process. I'll go over the answers again.

The student now provides some overt evidence of reasoning. The patient findings have been interpreted and a possible explanation (arterial problems, arterial insufficiency) of the patient's signs has been provided. The examinee then turns to the possible answers.

This initial activity of reading the entire item appears to be one of familiarization and usually serves as a forerunner to a more in depth analysis of the item which is described in the next student-item interaction.

A2. *Reads stem, searches for answer among alternatives, answer is triggered by alternative, eliminates remaining alternatives.* In this activity, overt indication of information processing may be revealed as the stem is being read. Key features may be defined, described, or expressed in the examinee's own words. The examinee does not generate an answer, but a broad answer space, for example, arterial problems, may be stated. The answer space represents an approximation to the best or correct answer. Alternatives are processed individually and an explanation is usually provided as to why each alternative may or may not be a plausible answer.

In the following example, the student processes the alternatives in a sequential manner, beginning with the first one. For the purposes of this study, the order in which alternatives were processed was not considered as a relevant factor.

> I'm reading the question now. OK, this is a 55 year old man and had a MI one week ago and now has severe pain in the left leg. The leg is cool, pale, and pulseless. The most likely diagnosis is and looking at the first choice here which says deep vein thrombosis. Deep vein thrombosis is less likely because usually you see quite a bit of swelling and it doesn't mention any swelling here. So I'm going to the second one, ruptured left iliac aneurysm. If you have a ruptured aneurysm - now I'm thinking how high the pain is and where exactly tne pain is in the leg. It could be in the area of the iliac artery, I'm not sure and if you have a ruptured aneurysm you probably don't have any pulse distal to that, so that could be. Arterial embolism, I don't think that one is that likely because it has to be a really big emboli. AV fistula, that one somehow I don't think is likely because you have a fistula, you have shunting of blood and it won't cause such a severe pain, I suppose, if you don't have any swelling or anything. Arterial thrombosis, again, well I suppose it could happen. Because of the cool, pale, and pulselessness of the leg, sounds like you have a decrease in arterial supply to the leg, so the first one would probably not be the answer. Second one, I would expect a lot of hemorrhage and swelling for the second one, so it's not likely. So I would choose between 3 and 5 and I think 5 will be more likely. Oh, wait a sec, this guy had an MI before so I suppose he can have an emboli that originate from the heart and goes down to the leg but again, has to be a really big one. So because he's 55 year old too and he

> will probably have a lot of arthrosclerosis for developing a MI, I would probably pick number 5 as the best.

A3. *Reads stem, generates answer, searches for answer, alternative triggers different answer, discards initial answer and eliminates other alternatives.* This move is similar to the activities described in activity 2 above. However, during the processing of the stem an answer is generated. In searching the alternatives for a match to the generated answer, an alternative triggers another answer resulting in the rejection of the initially generated answer. Here is an example:

> OK, so it's a 55 year old man, one week after infarct and I guess it's a sudden onset of severe left leg pain. It's cool, pale and pulseless. So I mean without looking at it I'd probably say he had a, could be a DVT but I'll look at the choices. OK, so DVT is there. A ruptured left iliac aneurysm, yah, that could be it but I mean he might have picked up things before. Arterial embolus, yah, actually that's more likely because after MI there's a risk of throwing off clots and stuff. AV fistula, no. Arterial thrombosis, yah, that could be that too but I'd say arterial embolus more because after MI if he's had any sort of arrhythmias or anything which I might be assuming too much but I'd say 3.

A4. *Reads stem, generates answer, searches, selects generated answer, scans or ignores remaining alternatives.* This is a generate, search for match, and select activity. Once an alternative is selected as the correct answer, the remaining alternatives may be scanned or ignored. No explanations are forwarded for their unsuitability. The following think aloud protocol illustrates this activity.

> OK, I'm reading the question. I'm noting that it's one week after his MI, anterior MI doesn't really help me that much at this point. Severe pain in his leg. Cool, pale and pulseless leg, that fact that it's pulseless makes me think of something that, like an arterial embolism so I'm looking at the answers and I see arterial embolism there, so, and that's definitely possible after an anterior MI and so I'm just looking at the other answers to see what there is but I'm going with my feeling that it's arterial embolism, so I'm choosing that one.

A5. *Reads stem, generates answer or answer space, searches, selects, eliminates alternatives, provides rationale for most of the eliminated alternatives.* This move differs from the activities presented in 4 above in several ways. First, an answer space, for example, a class of drugs rather than a specific drug from that class may be generated. Second, reasons are provided for the eliminated alternatives.

An example of a protocol in which a student generates an answer which matches one of the alternatives and then proceeds to provide reasons why the other alternatives are incorrect appears below.

> Alright, we have a middle-aged man who has had infarction and pain in the leg. The leg is cool, pale, pulseless. The physical signs suggest an embolism. DVT is not, he would have a pulse. Ruptured left iliac aneurysm is no relation to infarction. Arterial embolism

> is quite likely as he had a heart attack. AV fistula - no reason for it. Arterial thrombosis - no mention of previous disease so I would have to say number three - arterial embolism.

In summary, when examinees first encounter a multiple choice item, they may exhibit one or more of several broad activities. These global item-related moves deal primarily with whether the examinees generate an answer prior to reading the alternatives and the actions taken by the examinees with respect to the alternatives.

From the above examples of students' think aloud protocols, it is evident that examinees deal with alternatives in a variety of ways. The next section identifies sixteen moves that relate to the disposition of the alternatives in the 30 multiple choice items.

B. Disposition of Alternatives

B1. *Reads an alternative and makes no comment.* This activity is characterized by the examinee either reading the alternative or stating that the alternative is being read. The examinee provides no information as to the suitability of the alternative as an answer.

B2. *Reads an alternative and states "No".* The examinee simply states that the alternative is not the answer. No explanation is provided for why the alternative is an unsuitable answer.

B3. *Reads an alternative and states "Not sure" or "I don't know".* The examinee reads the alternative and responds with either "not sure" or I don't know." It would appear that these responses generally reflect a lack of knowledge or an uncertainty about the alternative as a possible answer.

B4. *Scans alternatives.* This activity occurs after an alternative has been selected as an answer and the examinee states that the remaining options are being scanned. The examinee does not make any statements about the options that are being scanned. The scanning activity appears to be a confirmation that the selected answer is the best answer.

B5. *Guesses an alternative as correct answer.* For some items, all alternatives or a subset of alternatives appear equally plausible and an alternative is selected by the examinee stating "I guess."

B6. *Ignores an alternative.* For this activity, no overt evidence is provided to suggest that the examinee has processed the alternative.

B7. *Eliminates an alternative on basis of delimiting factor.* Alternatives may be discarded on the basis of delimiting factors such as age, gender, or ethnicity. For example, an examinee states "I doubt its rheumatoid arthritis because of his age..." or "It could be ankylosing spondylitis but that's more, again I think it affects females more often than males."

B8. *Eliminates an alternative on basis of predisposing factor.* In this activity, alternatives may be eliminated because of the association between risk factors and diseases. For example, carcinoma of the lung is more common in smokers than non-smokers.

B9. *Eliminates an alternative using oppositions.* The list of oppositions in medicine is large (Lemieux & Bordage, 1986). Patients may present with "acute" or "chronic" conditions; diseases may be "acquired" or "congenital" and their presentation may be "apparent" or "insidious"; symptoms may be of "gradual" or "sudden" onset and may have appeared "recently" or "long time ago" represent some of the oppositions. It should be noted that the pairs of oppositions are not exclusive to medicine.

B10. *Eliminates an alternative on basis of similarities.* In some items, examinees considered two alternatives to be so similar that choosing one as an answer would imply that the other was correct as well. For such items both alternatives would be discarded because of similarity.

B11. *Eliminates an alternative using category or class.* As with oppositions, many aspects of medicine can be classified. Such systems of classification include etiologic, anatomic, or pathophysiologic organizations of medical knowledge. Each of these could be further sub-classified. For example, anatomic can have categories such as cardiovascular, respiratory, and musculoskeletal. In eliminating alternatives, examinees state that the patient's presentation suggests a heart problem rather than a lung problem.

B12. *Eliminates an alternative on the basis of priority.* Several activities characterize this move. For a set of alternatives where an order or sequence of actions is suggested, an alternative may be discarded because it does not represent the first action. For other sets of options, alternatives may be eliminated because they are inappropriate, less appropriate, unimportant, or less important.

B13. *Eliminates an alternative in terms of likelihood.* After reading an alternative examinee states that the answer is "unlikely", "rare", or "less likely than " another answer.

B14. *Eliminates an alternative because it does not match the generated answer or the answer space.* For those items in which an answer or an answer space was generated, examinees may eliminate an alternative by stating that it does not match the answer or it does not belong to the answer space.

B15. *Eliminates an alternative because it does not match the data provided in the stem.* For items in which patient information was presented and the task centered on choice of either diagnoses, laboratory tests, or treatment and management, students eliminated alternatives because the information in the alternatives was not congruent with the data presented in the stem. For example, "I'm looking at the answers and number 5, I don't think its multiple sclerosis, it's not typical presentation of multiple sclerosis. Otitis media, I'm looking at number 3, I don't think that's likely, the symptoms don't really correspond to that." In the example above, the student eliminates alternatives because the description in the stem does not match the diagnosis of either multiple sclerosis or otitis media. It should be noted that the options are eliminated from a global rather than a specific perspective. That is, the examinee makes reference to an overall picture of the diagnosis as opposed to making reference to specific signs and/or symptoms associated with the diagnosis.

B16. *Eliminates an alternative on the basis of association.* In contrast to the elimination of alternatives from the perspective of a global picture, students sometimes eliminated an alternative by elaborating on specific pieces of clinical information presented in the item. The student then states how the association, for example, between the clinical data presented in the stem and the alternative aids in eliminating the alternative. The implication in this type of activity is that students encode medical concepts and procedures as "if - then" statements and that these statements of relationships may be used to rule in or rule out answers. The following examples show how the association between concepts and alternatives is used to eliminate alternatives.

> The leg being cool, pale and pulseless, I can tell that there must be something blocking the arterial circulation because if it's a venous blockage you would have a pulse and the leg would be warm.
>
> Deep vein thrombosis is usually cool and pale ... is usually hot and red and he can have a pulse.
>
> OK, then I would look back at the question - cool, pale, and pulseless. Is that really associated with DVT ... um OK, DVT usually they are warm, edematous, tenderness to palpation.

In each of the examples, the alternative, "deep vein thrombosis", is eliminated using reasoning which takes the form "If warm, red and pulse, then consider deep vein thrombosis as an answer but the leg is cool, pale and pulseless and so deep vein thrombosis is not the answer." In eliminating the alternatives, the negation may be used as well. For example, "if cool, pale and pulseless, then not deep vein thrombosis."

Several aspects should be noted about stating that the alternatives are eliminated using the "if - then" framework. First, examinees do not necessarily use the words "if" and "then" in eliminating an alternative; they may say "I don't associate deep venous thrombosis with a cool, pale and pulseless leg." Second, an important distinction in association and the condition-action "if - then" rule is that the decision is based on clinical information that, for example, characterizes a disease process. This distinction is made because it is possible to cast other alternative elimination behaviors in the "if - then" format. For example, alternatives which are eliminated on the basis of delimiting or predisposing factors, oppositions, similarities, or category may be put in the "if - then" form.

The above list of behaviors which relates to the disposition of alternatives provides an overview of the moves the students used when dealing specifically with the options in the test items. The list is not exhaustive and the students may use activities which are not included in the list.

The final category of activities lists those behaviors that are identified with successful problem solving.

C. Successful Problem Solving Activities

The activities in this category have been identified by researchers investigating the nature of expertise and expert problem solving in fields such as physics, genetics, and medicine. Researchers such as Patel, Groen, and Arocha (1990) and Smith (1991) advocate that expert problem solving is a subset of successful problem solving because novices can exhibit problem solving activities which are similar to those of the experts. The following list of four activities is based on the literature and may be thought of as a kind of generic list of successful problem solving activities.

C1. *Restates information presented in the item in own words.* This activity is characterized by the examinee presenting a summary of the information contained in the item. Here are several examples.

> So go back, middle-aged man, anterior myocardial infarct one week ago and now he's got pain in his left leg.
>
> I'm reading the stem, it's a question that's asking about a diagnosis and so they're looking at a patient post myocardial infarct one week so right now I'm thinking of sorting in my head the complications of an MI and then his age and he has pain in his left leg.

C2. *Focusses on key features and defines or redescribes in a different manner.* This behavior usually occurs conjointly with activity C1 (restates information in own words) and centres on relevant pieces of information presented in the stem. Depending upon the item, activity C2 may serve as a precusor to moves C3 and C4. In the following example,

> Before I'm looking at the answers, the symptoms of pale, cool and pulselessness suggests an arterial lesion likely an obstruction of the artery to the left leg and if I go to my answers.,

the student focusses on the leg description and associates the salient features of a leg which is cool, pale, and pulseless with evidence for an arterial lesion.

C3. *Generates answer or answer space using the information presented in the stem.* This category along with the next one represents a somewhat superficial attempt to classify the processing of information presented in the stem as an indicator of forward reasoning. Because the verbal transcripts obtained in this study were not subjected to propositional analysis, evidence for forward reasoning was determined by whether examinees, on the basis of the information provided in the stem, generated responses which were either identical to the alternatives (e.g. "...makes me think of something that, like an arterial embolism ...", "... first thing that comes to my head is a DVT ... ") or similar to the alternatives (e.g. "... there must be something blocking the arterial circulation ...", "... sounds like an arterial problem ..."). Activity C3 captures the "generates answer" component of activities A3, A4, and A5.

C4. *Activates hypothesis using the information presented in the stem.* This activity is closely related to activities C2 and C3 above. Depending upon the item, the responses generated by the examinees upon reading the stem, may represent an intermediate step in selecting an answer. For example, in an item requesting treatment for a patient presenting with a set of signs and symptoms, examinees may generate diagnoses which would explain the signs and symptoms, although the item did not request a diagnosis directly. Examinees may or may not generate an answer to the specific question of treatment.

It should be noted that moves C3 and C4 as conceptualized above reflect a student's ability either to generate an answer which may or may not match one of the alternatives or to activate a hypothesis which would represent an intermediate phase of reasoning. Forward reasoning is indicated if the student takes the data presented in the stem and then on basis of this information produces an answer or a diagnosis. It is also possible to think of forward reasoning as deducing new data from existing data, that is, all newly generated data by the student are based on the original facts. In contrast, backward reasoning is characterized by introducing new facts to account for the original data. The following segment of a protocol shows forward reasoning.

> O.K., let's read the item. The most likely physical finding to be noted in a patient during an attack of angina on effort, O.K., so this is stable angina, angina on effort. The most likely finding to be noted during an attack of angina on effort is a: fourth heart sound, cardiology, an attack of angina on effort. So if there is angina on effort, the myocardium is going to be a little bit ischemic, may have loss of compliance.

The student's statement that this is a case of stable angina and that during stable angina the myocardium has a lack of oxygen due to inadequate perfusion (ischemia) and may be non-compliant would be an indication that the student is generating new material from existing data presented in the stem of the item. Viewed in this manner, combining moves C2, C3, and C4 may provide evidence of forward chaining.

In summary, the above list of moves was developed partially from the activities students displayed in their interactions with the multiple choice items and partially from the literature on expertise in medicine and successful problem solving. The list consists of five global activities that relate to the overall approach to the item, sixteen activities that relate to the disposition of alternatives, and four activities that characterize successful problem solving.

Following the development of the list of examinee-item interaction activities, the transcripts were coded according to the behaviors described in the three categories. A think aloud protocol for item 1, which follows, will serve as an example of the coding.

> One week after an anterior myocardial infarction a 55-year-old man complains of severe pain in the left leg. The leg is cool, pale and pulseless. The most likely diagnosis is:
>
> 1. Deep venous thrombosis.
> 2. Ruptured left iliac aneurysm.
> * 3. Arterial embolism.
> 4. A-V fistula.
> 5. Arterial thrombosis.

> Ok, just looking at the question the first diagnosis that comes to mind is that he has thrown off an emboli from the infarct and it's blocking the femoral artery probably. When you say the leg is cool, pale, pulseless it would be nice to have the location of the pulses, which pulse is absent, is it anterior tibial or posterior tibial, popliteal or dorsalis pedis. Now I'll look at the answers. Deep venous thrombosis doesn't work because it is not right - that sort of clinical, the clinical presentation would be different. Ruptured left iliac aneurysm is probably not very likely because then what does the infarct have to do with it. Arterial embolism is the one I said. A-V fistula, what is the point of giving information about the myocardial infarction if it's an A.V. fistula. Arterial thrombosis, that is not very likely in the leg. I'll say 3, arterial embolism, on that basis.

With respect to the overall approach to this item, this examinee's protocol was coded as A5 (*Reads stem, generates answer, searches, selects, eliminates alternatives, provides rationale for most of the eliminated alternatives*). For the disposition of alternatives, option 1 was coded as B15 (*Eliminates an alternative because it does not match the patient description provided in the stem*); alternatives 2 and 4 were coded as B16 (*Eliminates an alternative on the basis of association*), and alternative 5 was coded as B13 (*Eliminates an alternative in terms of likelihood*). In terms of successful problem solving activities, the above protocol was coded as C1 (*Restates information presented in the item in own words*) and C3 (*Generates answer or answer space using only the information presented in the stem*).

At the conclusion of the coding step, each think aloud protocol had a series of alphanumeric codes (e.g. A5, B15, B16, B16, B13, C1, C3) which reflected the activities the student in the above protocol used in answering the first item. In the set of 30 items, items 5 and 6 were based on the same patient scenario and because the responses to item 6 were dependent upon the responses given to item 5, item 6 was deleted from the analysis. One student answered items 5 and 6 as a unit. This student's protocol for item 5 was not part of the analysis. Item 8 was misread by one student and as a result this student's protocol was deleted. For item 30, a student's treatment of two alternatives was lost during the changing of tape sides. The responses to the two alternatives for this student were not included in the analysis.

The second phase of the initial analysis consisted of tabulating the student-item interaction activities. These tabulations appear in Appendix VII of which a portion is reproduced in Table 5. With reference to Table 5, the tabulations for the item-related and the problem solving activities represent the number of students displaying that activity. For the alternatives, the figures represent the number of times, across alternatives and students, the activity was offered as a reason for rejecting an alternative. For item 1, four students read the stem and the alternatives as a unit, eight students read the stem and then proceeded to read the alternatives and search for the answer, seven students upon reading the stem, generated an answer but changed the answer upon reading the alternatives, five students generated an answer from reading the stem and offered very few or no reasons for the disposition of the alternatives, and finally twenty students generated an answer and offered reasons for the selection and rejection of alternatives.

Students used a variety of moves when dealing with the alternatives in the items. For item 1, the activity *"Reads an alternative and makes no comment."* (Move B1) was used twelve times. A review of the students' protocols showed that the activity, *"Reads an alternative and makes no comment."* was used twice for alternative 2, once for alternative 3 and five and four times for alternatives 4 and 5 respectively. The most frequently appearing activities for disposing alternatives were "*Eliminates an alternative in terms of likelihood*", *"Eliminates an alternative because it does not match the patient description provided in the stem"* and "*Eliminates an alternative on the basis of association*".

Table 5

Frequencies of Move Use for Items 1 to 5

	ITEMS				
MOVES	1	2	3	4	5
A. ITEM RELATED					
1. Reads item	4	15	1	3	-
2. Read and search	8	39	33	27	14
3. Generate - trigger	7	-	-	1	3
4. Generate - no reason	5	1	-	10	7
5. Generate - reason	20	-	7	2	15
B. ALTERNATIVES					
1. No comment	12	32	21	19	9
2. No	11	29	7	6	8
3. Do not know	3	4	8	6	3
4. Scans	4	-	-	-	-
5. Guesses	-	4	-	4	-
6. Ignores	2	-	1	65	13
7. Delimiting	-	-	-	-	16
8. Predisposing	2	1	-	-	1
9. Oppositions	4	2	5	-	2
10. Similarities	-	-	-	-	-
11. Category	6	20	1	35	1
12. Priority	-	-	-	-	-
13. Likelihood	39	8	9	18	9
14. Answer	10	-	-	-	10
15. Description	24	17	47	6	37
16. Association	43	47	61	5	47
C. PROBLEM SOLVING					
1. Restates	32	23	30	5	14
2. Redescribes	26	19	32	8	10
3. Generates answer-stem	32	1	7	13	25
4. Activates hypothesis	-	-	39	-	-

Considering the problem solving activities, 32 of the 40 students restated the information presented in the item (Move C1), 26 students defined or redescribed the key features presented in the item (Move C2), and 32 students generated an answer on the information presented in the stem and without reading the alternatives (Move C3).

A further review of the tabulations for the items presented in Table 5 and the remainder of the items in Appendix VII reveals that students used a variety of item-related activities in their interactions and that these activities differed across items. For items 1 and 5, more than half of the students generated answers without considering the alternatives (C3=A3+A4+A5). In contrast, for items 2, 3, and 4, students depended upon the alternatives to trigger an answer (Move A2), although 13 of the 40 students did generate an answer (Move C3) for item 4 before reading the alternatives. Furthermore, different students provided different explanations for the rejection of the same alternatives and displayed some of the characteristics of successful problem solving. For some items, students displayed behaviors which were intermediary to answering the item. As an example, item 3 presented a scenario of a patient with signs and symptoms suggestive of congestive heart failure. The item asked what the patient's x-ray findings would show. In processing the stem, six students generated an answer related to x-ray findings. Thirty-nine students activated diagnoses associated with heart disease.

Reliability of the coding of the entire set of think aloud protocols was not addressed in a formal manner. However, once the list of moves was finalized and the coding completed, the classification of each student-item engagement was repeated several months later. Overall, few changes were made to the classifications with respect to the item-related and successful problem solving activities. In terms of the activities related to the disposition of alternatives, problems in classification were encountered with activities B15 (*Eliminates an alternative because it does not match the data provided in the stem.*) and B16 (*Eliminates an alternative on the basis of association.*). On rereading the protocols, alternatives which were coded as B15 were recoded as B16 and vice-versa. In terms of the purposes of the study, the ambivalence in the coding of alternatives using activities B15 or B16 may be insignificant. Nevertheless, future work in terms of identifying activities and strategies used by students responding to multiple choice items should provide formal evidence for intra- and inter- rater agreement for coding.

Further Analysis

Since one of the goals of the study is to identify and describe the strategies students use when responding to multiple choice items as well as to determine whether strategy commonalities are exhibited across items, parameters such as item difficulty, system, and task (Know, Comprehend, Diagnose, Test, and Manage) were used to group the items. Using these parameters produced item groupings which did not aid in making meaningful comparisons. That is, students' approaches to answering items did not appear to be related either to the item's difficulty, task, or the system in which the item was grounded. Although the task presented in each stem may have helped focus the item, the initial analysis of the student-item interactions showed considerable variability in students' encounters with items. It would appear that a student's approach to an item is some complex combination of the semantic features of the item, the student's declarative and procedural knowledge, and the student's metacognitive skills. Because no discernable patterns of responses could be identified using parameters such as item difficulty, system, or task, several moves were selected from the list of test-taking activities presented earlier in this chapter as a basis for grouping items. These moves related to whether students generated an answer and whether students activated a hypothesis. Using these two criteria produced three groups of items.

Group One consisted of those items in which "few or no answers were generated and no hypotheses were advanced". Items 2, 8, 10, 25, 26, 29, and 30 comprised the first group. A scan of these items would suggest that the items tap declarative knowledge.

The second group, consisting of items 3, 7, 12, 13, 17, 18, and 19 were items for which "few or no answers were generated but hypotheses were activated". A glance at these items suggests that the items bear some resemblance to the paper cases used in the investigation of the clinical reasoning construct by researchers such as Lemieux and Bordage (1992), Ramsden, Whelan and Cooper (1989), and Patel and Groen (1986). That is, the stems contain clinical presentations of patients with medical problems, although, the amount of information presented may be less than that appearing in the paper cases.

The third group consisted of the remaining items in which students "generated answers but few or no hypotheses were activated". This is a mixed group of items; some items are based on clinical presentations of patients, much like the items in Group Two, while other items are more like the items in Group One.

In determining these groupings, "few" was arbitrarily defined as less than seven instances of the activity. Appendix VIII presents the tabulations for the regrouped items.

The activities used by the students in each of the three groups were examined in more detail by considering the list of moves identified in the initial analysis and the frequency with which the moves were used. In addition, flow charts were developed to illustrate the global patterns used by the students to answer the items in each group. The flow charts are based on the five moves identified in the item-related activities (A1-A5) and the two moves associated with the generation of answers and activation of hypotheses (C3 and C4) identified as part of successful problem solving activities. The use of flow charts is based on the suggestion made by Afflerbach and Johnston (1984). According to Afflerbach and Johnston, one of the problems that arises from the classification and description of the activities obtained from think aloud protocols is that the sequencing of activities may be lost. According to them, flow charts may be developed to determine the frequencies and sequences of the activities and to illustrate the solution paths used.

This chapter dealt with the selection of examinees and items, procedure, and initial analysis. The initial analysis of the think aloud protocols identified a number of different activities or moves. From the students' global approaches to each item as a whole and from the students' activities involved in the disposition of the alternatives, five item-related and sixteen disposition of alternatives activities were identified. From the moves associated with processing the entire item and the literature on expertise and problem solving, four successful problem solving moves were inferred. This list of twenty-five activities serves as a preliminary framework for coding the think aloud protocols of the student-item engagements.

The next chapter presents the results of the strategy analysis.

IV. RESULTS AND DISCUSSION

Several different pieces of information are presented and discussed in this chapter. First, the responses of the forty students to the questionnaire which appears in Appendix V are reported. This is followed by examples of students' think aloud protocols showing the different activities that emerge during the student-item interactions for each of the three item groups. These activities or moves are combined into strategies and these strategies are examined in terms of the literature on expertise in medicine and successful problem solving.

Responses to the Questionnaire

At the conclusion of the interview, each student completed a questionnaire consisting of twenty-three items that requested opinion on 1) purpose of study, 2) instructions, 3) think aloud examples, 4) when thinking about the test items, 5) the exam, 6) exam content, and 7) the test items. Each of these aspects was represented by sets of adjective or description pairs and opinions were specified using a four-point scale. The percentage of students selecting each point of the adjective or description pairs for these components is presented below. Percentages summed across the four points do not necessarily equal 100% because of omissions. The first set is for purpose of study and instructions.

Purpose of study:

1.	explained clearly	90.0 : 7.5 : 2.5 : 0.0	lacked clarity
2.	informative	47.5 : 47.5: 0.0 : 0.0	uninformative

Instructions:

3.	clear	95.0 : 5.0 : 0.0 : 0.0	ambiguous
4.	complex	12.5 : 10.0 : 27.5 : 47.5	simple
5.	easy to follow	67.5 : 30.0 : 0.0 : 0.0	difficult to follow

With respect to the first two topics, the majority of the students felt that the purpose of the study was informative and explained clearly. The instructions to the subjects were clear, simple and easy to follow, however, nine (22.5%) students felt that the instructions were complex.

The responses of the students to the example verbal reports which were read prior to the interaction with the experimental items are as follows.

Think-aloud examples:

6.	useless	5.0 : 12.5 : 40.0 : 40.0	helpful
7.	influenced my thinking	7.5 : 37.5 : 27.5 : 27.5	did not influence my thinking
8.	influenced my responses	7.5 : 20.0 : 25.0 : 45.0	did not influence my responses

The majority of the students considered the examples as helpful. However, they were split on the influence of the examples on their thinking. About one quarter of the students stated that the examples influenced their responses. A similar pattern of response is obtained when students were questioned on their thinking about the experimental items.

When thinking about the test items:

9.	probing by the researcher interfered with my reasoning	0.0 : 10.0 : 22.5 : 65.0	probing by the researcher did NOT interfere with my reasoning
10.	my thinking and reasoning were typical of how I approach exams	62.5 : 25.0 : 12.5 : 0.0	my thinking and reasoning were NOT typical of how I approach exams
11.	I used strategies that I typically use on other exams	67.5 : 30.0 : 2.5 : 0.0	I did NOT use strategies that I typically use on exams
12.	I mirrored my thinking on the examples	25.0 : 30.0 : 25.0 : 20.0	I did NOT mirror my thinking on the examples provided

Slightly more than one-half of the students mirrored their thinking on the examples provided but the majority of them stated that their thinking, reasoning and approach was typical of other exams.

The remaining questions pertained to qualities of the exam, exam content, and items.

This exam was:

13.	difficult	7.5 : 75.0 : 12.5 : 2.5	easy
14.	testing essentials	15.0 : 67.5 : 15.0 : 2.5	testing obscurity
15.	out-dated	0.0 : 7.5 : 57.5 : 32.5	current
16.	fair	17.5 : 70.0 : 10.0 : 0.0	unfair
17.	like other exams	50.0 : 32.5 : 10.0 : 5.0	unlike other exams

The examination content represented:

18.	course or rotation objectives	37.5 : 45.0 : 12.5 : 2.5	no course or rotation objectives
19.	material taught in classes	27.5 : 65.0 : 7.5 : 0.0	material not taught in classes
20.	material presented on the ward	12.5 : 50.0 : 25.0 : 10.0	material not presented on the ward

The test items were:

21.	clear	12.5 : 70.0 : 15.0 : 2.5	ambiguous
22.	incomplete	2.5 : 25.0 : 72.5 : 0.0	complete
23.	poorly designed	0.0 : 22.5 : 65.0 : 12.5	properly designed

At least 80% of the students judged the exam to be difficult but fair and testing current and relevant information. The content tested represented course or rotation objectives and material taught in class although about one-third of the students felt that the exam was not representative of the content on the ward. The items appeared unambiguous, complete, and properly designed.

Examples of Think-Aloud Protocols

Results for each of the three item groups identified in the *Further Analysis* section of Chapter III will be presented in terms of the three categories of moves -- item related, disposition of alternatives, and successful problem solving activities. For some of the items, item related and successful problem solving activities will be considered jointly because of the overlap and similarity between the two categories. Reference is also made to the tabulations presented in Appendix VIII.

Group One Items

The first group consists of seven items (2, 8, 10, 25, 26, 29, and 30) for which the students generated few or no answers and for which no hypotheses were activated. In terms of student-item interactions and the activities used to answer the items, several salient features characterize the seven items. Considering the item related moves first, the two dominant activities exhibited across the seven items were A1 -- *Reads stem and all alternatives* and A2 -- *Reads stem, searches for answer among alternatives, answer triggered by alternative, eliminates remaining alternatives.* The number of times the entire item was read as a unit or where the stem was read first and the information presented in the stem was interpreted after which the alternatives were read as a unit (Move A1) ranged from three times for item 10 to twenty times for item 8. The use of activity A2 ranged from 34 to 39. For three of the items (2, 10, and 25), the two item related activities of A1 and A2 were used in conjunction with two successful problem solving moves C1 -- *Restates information presented in the stem in own words* and C2 -- *Focusses on key features and defines or redescribes in a different manner.* For the remaining four items (8, 26, 29, and 30), students used the two successful problem solving activities, C1 and C2 with less frequency.

To illustrate the use of moves A1, A2, C1, and C2, the protocols of two students responding to item 2 are presented below. The stem for item 2 reads,

> The most likely physical finding to be noted in a patient during an attack of angina on effort is a:
>
> * 1. Fourth heart sound.
> 2. Second sound of diminished intensity.
> 3. Mid-systolic murmur at the cardiac apex.
> 4. Transient aortic ejection click.
> 5. Third heart sound.

In the first example, the student began by reading the entire item (Move A1).

> The most likely physical finding to be noted in a patient during an attack of angina on effort is a: A - fourth heart sound, B - second sound of diminished intensity, C - mid systolic murmur at cardiac apex, 4 - transient aortic ejection click, 5 - third heart sound.

For this student reading the entire item may have served as an advance organizer, that is, a precusor to obtaining an initial idea of the information presented in the item and the task requested of the examinee. Having read the entire item, the student then proceeded to say:

> During an attack of angina so you think of the reasons the patient may have angina, just thinking of coronary artery insufficiency would be number one on my list, you could also get angina from an increased diastolic filling time, seeing as the heart is perfused only in ... diastole ... was a

> little mixed up there. The heart is perfused only during diastole, so if diastolic pressure is increased as well as the time being shortened, you are going to get a decreased amount of blood. So we're looking for the physical findings in this guy.

In the second part of the protocol, the student focussed on the key feature -- angina, generates one possible reason for angina, provided the pathophysiology for angina, and restated the task posed in the stem (Moves C1 and C2). This student's attempt at providing a reason (coronary artery insufficiency) for the angina and the student's explanation of why angina occurs may be taken as a rudimentary form of forward reasoning. It is rudimentary for two reasons. First, the student having read the stem and the alternatives as a unit should have a good idea of what the item requests. Second, the item requests a physical sign of angina on effort and this should trigger the student to think about cardiology and the heart. Nevertheless, in the above protocol the student built a description (although perhaps incomplete) of the etiology and pathophysiology of angina. In this segment of the protocol, the student used forward reasoning by generating new data from existing information. If backward reasoning were to be displayed, one would speculate that the student might begin by listing the systems, such as cardiovascular and respiratory, which have something to do with the regulation of oxygen. Within each system, a number of differential diagnoses such as asthma, chronic obstructive lung disease, angina pectoris, carcinoma, and myocardial infraction might be generated and the student would then go through a process eliminating those that were not connected with angina on effort. Examples of this type of reasoning did not appear in the think aloud protocols for item 2.

Continuing with item 2 and the first example of a think aloud protocol, this student incorporated a redescription of the problem, extracted the important components and represented the item in terms which gave personal meaning and understanding. Newell and Simon (1972) and Smith (1991) refer to this cluster of activities as the creation of an internal problem space. After internalizing the information presented in the item, the student went on to the alternatives.

> The fourth heart sound suggests increased diastolic filling pressure which would go with the decreased perfusion of the myocardium. The second heart sound of diminished intensity would suggest a problem with either the aortic or pulmonic valves. Mid systolic murmur at the cardiac apex would suggest problems with mitral regurgitation which could secondarily cause an increase in the left ventricle and a distolic filling pressure which could cause you angina. But you would notice this during not only during the acute attack but also during a period when he is asymptomatic. Transient aortic ejection click, again it's suggesting a problem with the aortic valve which could lead to increased left ventricular filling pressure. Third heart sound you would hear when the myocardium has decreased compliance. This could be due to left ventricular and diastolic filling pressure increase, also dilatation of the heart. So you wouldn't be able to hear a third sound, possibly you would hear a fourth heart sound. You could hear a mid systolic murmur at the cardiac apex and you may be able to hear that during the acute attack but seeing as they are wanting to

> know the physical findings during the attack and not previous to it, I wouldn't say it's a mid systolic murmur. For the same reason, I wouldn't say it's the aortic ejection click and for the same reason too, I wouldn't say it's the second sound of diminished intensity. Fourth heart sound, third heart sound, they are the more acutely occurring heart sounds. I would say because the fourth heart sound occurs more when the pressures increase, I would say number 1 over number 5.

Because an answer to the item was not expressed either upon reading the stem, reading the entire item, or after internalization, it is inferred that the student derived a response from processing the alternatives (Move A2). The manner in which students dealt with alternatives will be discussed later. However, in this example, the student provided a reason as to the suitability of each alternative as an answer. Alternatives were eliminated because the physical findings are not consistent with angina but more consistent with other diagnoses such as valvular problems and mitral regurgitation. Furthermore, the student searched for an answer that meets the underlying causes of angina which the student presented during the internalization of the problem. In a way, each of these alternatives served as a provisional hypothesis which the student discarded through backward reasoning. For this student, the process continued until a choice had to be made between two alternatives and the selection of answer was based on the student's explanation of the pathophysiology presented earlier in the think aloud protocol.

In the next example, the second student's overall approach to the same item was similar to that of the first student. Differences do arise in the way in which the information presented in the stem was handled and the manner in which the alternatives were treated. The student began with the stem.

> The most likely physical finding to be noted in a patient during an attack of angina on effort is ok so physical finding something that you pick up on physical exam as opposed to symptom, angina on effort. OK well here I don't really have anything that really strikes me right away although angina is due to decreased oxygen to the heart. So this one I would probably just look through the answers.

In the first part of the protocol, the student read the stem, noted that a sign (physical finding) as opposed to a symptom was required. The student indicated that a self-generated answer was not available, stated a cause of angina, and then stated that the answers will be surveyed. The student exhibited the successful problem solving activities C1 and C2. This student's initial internalization of the problem was based on the interpretation of the information presented in the stem only. The student then read the five alternatives as a unit (Move A1) and stated that unreasonable alternatives were to be eliminated. These activities appeared in the second part of the protocol.

> Fourth heart sound, second sound of diminished intensity, mid-systolic murmur at the cardiac apex, transient aortic ejection click, third heart sound. OK now I'm just going to rule out ones I don't think make sense. Second sound of diminished intensity doesn't have anything to do

> with angina as far as I know, mid-systolic murmur at the cardiac apex is the same kind of thing, transient aortic click, unlikely, so it's probably third heart sound or fourth heart sound. An attack of angina on effort I would say it's decreased contractility of the heart so the S3, I'm just trying to think if S3 is due to volume overload, one of them is due to volume overload and the other is due to pressure overload. As far as I can remember the pressure overload is related to ... hm, no let me think about this, pressure overload I believe is a fourth heart sound, so I would just put one.

Since the student was unable to generate an answer on reading the stem, reading the alternatives as a unit may be the student's way of creating an internal problem space and placing some boundaries on what is required by the item. Through a process of elimination, the student reduced the number of plausible answers to two choices from which one was selected and confirmed. Because the student did not generate an answer prior to reading the alternatives, the answer to the item was triggered by the alternatives (Move A2).

Although both students answered the item correctly and displayed the moves A1 -- *Reads stem and all alternatives*, A2 -- *Reads stem, searches for answer among alternatives, answer triggered by alternative, eliminates remaining alternatives*, C1 -- *Restates information presented in the stem in own words* and C2 -- *Focusses on key features and defines or redescribes in a different manner* several differences appear in their approaches to the item. The first student read the stem and the alternatives as a unit and then proceeded to redescribe the information contained in the stem. This was in contrast to the second student who did the redescription upon reading the stem only. The comprehensiveness of the redescription also differs. The first student provided a more in depth analysis of the cause of angina whereas the second student merely defines angina. It would appear from the protocols that the second student attempted to generate an answer but was unsuccessful whereas no such attempt was made by the first student. Finally, with respect to the alternatives, the second student dismissed alternatives 2, 3, and 4 as being unrelated to and unlikely physical findings of angina. The first student also eliminated these three alternatives but the reasoning was quite different: they were eliminated because they are physical findings associated with conditions other than angina. In addition, the student reinforced the deletion of these alternatives with reference to the underlying pathophysiology for angina which the student developed in the redescription of the item. The second student resorted to the pathophysiology only in deciding whether the correct answer was the fourth or third heart sound.

With respect to the item-related activities, the pattern of response exhibited by most of the students for the seven items in Group One is one of reading the item and searching the alternatives for an acceptable answer. The activities associated with generating an answer prior to reading the alternatives were used infrequently. Item related activities A3, A4, and A5 were exhibited a total of seventeen times from a maximum possible of two hundred and seventy nine. Each of these seventeen instances is also an indicator of move C3 -- *Generates answer or answer space using the information presented in the stem.* Of the seventeen instances where an answer was generated on the basis of the information contained in the stem, nine exhibited move A3 -- *Reads stem, generates answer, searches for answer, alternative triggers different answer, discards initial answer and eliminates other alternatives.* In this activity a generated answer is changed to one matching an alternative. To illustrate this, a student's response to the following item (Item 29) will be used.

> Generalized pruritus is most commonly due to:
> 1. Hodgkin's disease.
> 2. Polycythemia.
> 3. Hyperthyroidism.
> 4. Chronic renal failure.
> * 5. Xerosis (dry skin).

The student's think aloud protocol begins as follows.

> Generalized pruritis is most commonly due to, pruritis, itching, generalized pruritis I would think maybe a drug reaction would be the most common because ... um, well, what else would give you; allergy or something.

Up to this point the student has read the stem, defined pruritis as itching, and has generated the answers, *drug reaction* and *allergy*. The student then continued by reading the five alternatives -- "Hodgkin's, Polycythemia, Hyperthyroid, Chronic liver failure, Dry skin." Since the student was unable to find a match between the generated answers and the alternatives, the answers *drug reaction* and *allergy* were replaced with the fifth alternative -- *dry skin.* The student proceeded to deal with the confirmation of the selected answer and the remaining alternatives in the following fashion.

> So being common, generalized dry skin can be, I would pick number five just because I don't think the other ones are as common as dry skin. Chronic renal failure isn't, Polycythemia isn't, Hodgkin's certainly isn't, Hyperthyroid is probably the most common of the other ones but dry skin is more common. It seems too easy for a medicine question, it seems too obvious, generalized pruritis, I stay with number five.

Of the remaining eight instances in which an answer or an answer class was generated, activity A4 -- *Reads stem, generates answer, searches, selects generated answer, scans or ignores remaining alternatives* -- was used three times. The following protocol for the item requesting a physical finding of angina shows the generation of an answer prior to reading the alternatives. After selecting the generated answer, most of the remaining alternatives are either ignored or read without further comment.

> So during an attack of angina, the physical finding. I think, without looking at the answers, I think third heart sound is pathological in congestive heart failure or even angina. OK, but let's see. OK, fourth heart sound, Second sound of diminished intensity. I think it's number 5, third heart sound, just from what I've read if I remember correctly and ... OK looking at the rest. Fourth heart sound, I don't think so, I'm not exactly too sure so I'll just ignore that. I really don't think it's the rest because just from what I remember I'm not too sure but third heart sound for some reason sticks in my mind.

In comparison to the use of activity A4 as shown in the above protocol, activity A5 -- *Reads stem, generates answer or answer space, searches, selects, eliminates alternatives, provides rationale for most of the eliminated alternatives* was used in five student-item interactions. Out of the seven items that comprise Group One, move A5 was used only with items 29 and 30. Examples of two students' verbalizations for the items follow. The first is for item 29 which was presented earlier and used as an example for a student using move A3.

> Generalized pruritis is most commonly due to itchy skin or dry skin from what I know, so I'll just look at the answers. Hodgkin's disease, that's not that common. Polycythemia, it is but that's not that common. Hyperthyroidism, that's not the most common cause from what I know. Chronic renal failure, uremia can give you pruritis but that's not the most common. So dry skin, I pick number 5.

For this item, the student generated an answer -- *itchy skin or dry skin* and then proceeded to the alternatives searching for a match. Alternatives were discarded on the basis of whether the disease or condition is a common cause of generalized pruritis.

The second item in this set for which students used move A5 was item 30. This item is presented below.

> The irritable bowel syndrome in adults ("irritable colon") is a diagnosis of exclusion. However, when this diagnosis is finally made you should:
>
> 1. Tell the patient the symptoms are always due to emotional stress.
> 2. Tell the patient to routinely take tranquilizers when symptoms flare.
> 3. Tell the patient to return for a complete reevaluation (x- rays, blood work, etc.) in three months.
>
> * 4. Counsel the patient and prescribe metamucil and bran.
>
> 5. Counsel the patient and prescribe Lomotil and Kaopectate.

For this item, the student read the stem and then generated three possible approaches for dealing with patients with irritable bowel.

> Irritable bowel, irritable colon is a diagnosis of exclusion, however, when the diagnosis is finally made you should; OK, well I just remember a lecture on irritable bowel we got and I'm trying to figure out what it said because it's such a common problem that I think, I didn't pay a lot of attention to it. I think the first one is education, second one is I think some mild laxatives and things like that, and the third one if those don't work is co-allergen agent, so I'd tell ... .

The answers, which the student appeared to be retrieving from a lecture, are general in the sense that the educational intervention and the choice of laxatives and drugs are all non-specific and therefore may be thought of as answer classes. The student then went on to the alternatives and dealt with them in the same order as they were presented in the item.

> Tell the patient the symptoms are always due to emotional stress, that's not true because I know that there is some physiological reason for that, so that's obviously wrong. Any ways, the thing is that even if you weren't a doctor and you knew nothing about medicine you'd know that answer would be wrong. Tell the patient to routinely take tranquilizers when symptoms flare. I think if you knew anything about medicine you'd know that was wrong answer too because you never say routinely take tranquilizers. Tell the patient to return for a complete re-evaluation in three months, well that's wrong because it you've made the diagnosis you know the diagnosis is exclusion so what's the point in coming back for a re-evaluation. Counsel the patient and prescribe Metamucil and bran, that could be true. Counsel the patient and prescribe Lomotil and Kaopectate, alright it's between 4 and 5 and the thing is even if you didn't know anything about irritable bowel you'd know it would be either 4 or 5. You know, I can't really see you prescribing Lomotil on a regular basis so I'd go with Metamucil and bran, number 4. The thing is bran can be given through diets any ways.

For each alternative a reason was provided for its appropriateness as a suitable answer. From the set of alternatives, the choice of plausible answers was reduced to two; drug therapy was eliminated as suitable management and the student selected the keyed answer.

Thus far, item-related and successful problem solving activities have been considered for Group One items. In terms of item-related activities, students used the moves A1 -- *Reads stem and all alternatives* and A2 -- *Reads stem, searches for answer among alternatives, answer triggered by alternative, eliminates remaining alternatives* for the majority of the items in this set. Activities associated with the generation of an answer, answer class, or answer space on reading the stem only (Moves A3, A4, A5, and C3) were used infrequently. Successful problem solving activities (Moves C1 and C2) were used with items 2, 10, and 25 more frequently than with the remaining four items. Reasons for this will be discussed later. The successful problem solving activity C4 -- *Activates hypothesis using the information presented in the stem.* was not observed in the think aloud protocols for the seven items. In terms of the description of move C4, it is necessary to reiterate that the activity represents an intermediate step in responding to an item. Activity C4 refers specifically to the situation where the student generates a diagnosis but the task set by the item is other than diagnosis. For example, the item may request appropriate treatment, next course of management, or laboratory investigations.

Dispensing of Alternatives

A review of the think aloud protocols presented thus far, shows that students use a variety of reasons for dispensing with the incorrect alternatives. In the next section the activities associated with the disposition of alternatives for Group One items will be considered.

A survey of the tabulations presented in Appendix VIII shows that of the sixteen moves related to the processing of alternatives, some are used more frequently than others. The frequency of move use for the alternatives for the Group One items is reproduced in Table 6. Several aspects of Table 6 bear further note. First, frequency counts were obtained for the entire set of alternatives associated with an item. For example, for item 2, move B1 -- *Reads an alternative and makes no comment* was observed 32 times.

Table 6

Frequency of Move Use for Alternatives of Group One Items

		Items							
	MOVE	2	8	10	25	26	29	30	Sum
1.	No comment	32	7	11	18	16	14	8	106
2.	No	29	1	5	30	51	20	62	198
3.	Do not know	4	1	-	27	16	13	1	62
4.	Scans	-	-	3	3	-	-	-	6
5.	Guesses	4	2	4	-	-	-	-	10
6.	Ignores	-	4	7	1	-	8	-	20
7.	Delimiting	-	1	-	4	-	-	-	5
8.	Predisposing	1	-	-	-	-	-	-	1
9.	Oppositions	2	140	-	-	8	49	1	200
10.	Similarities	-	-	12	-	-	-	2	14
11.	Category	20	-	-	-	13	-	-	33
12.	Priority	-	-	-	-	7	-	67	74
13.	Likelihood	8	2	2	5	27	47	-	91
14.	Answer	-	-	14	-	-	-	-	14
15.	Description	17	-	106	10	4	1	5	143
16.	Association	47	-	-	62	18	8	12	147
	TOTAL	160	156	160	160	160	160	158	1114

Second, summing the number of times the activity appears across the seven items indicates the overall frequency with which each move is used. From Table 6, it is observed that move B8 --

Eliminates an alternative on basis of predisposing factor was used only once while activity B9 -- *Eliminates an alternative using oppositions* was used 200 times. Third, some activities were used across all or the majority of the items, for example, moves B1, B2, B3, B9, B13, B15, and B16 while other moves such as B4, B7, B8, and B14 were only used for a subset of the items. Finally, for some of the items, the elimination of alternatives is dominated by one or two activities. For item 8, students used move B9 -- *Eliminates an alternative using oppositions* for the majority of the alternatives while move B15 -- *Eliminates an alternative because it does not match the data provided in the stem* dominates the manner in which alternatives were treated in item 10.

To show how students used the sixteen moves associated with the disposition of alternatives, specific examples of think aloud protocols will be presented. The examples will concentrate on the more frequently used moves. However, in presenting the results for the three groups of items, the use of all the moves associated with the disposition of alternatives will be presented.

From Table 6, the most frequently used activities are B1 -- *Reads an alternative and makes no comment*, B2 -- *Reads an alternative and states "No",* B9 -- *Eliminates an alternative using oppositions,* B15 -- *Eliminates an alternative because it does not match the data provided in the stem*, and B16 -- *Eliminates an alternative on the basis of association.* The number of times each of these strategies is used ranges from 106 to 200. Strategies B1 and B2 appear for each item and account for about 27% of the activities associated with disposition of alternatives. Item 25, which is reproduced below, will be used to show the use of move B1.

> A diagnostically helpful ophthalmic finding in lupus erythematosus is:
> * 1. Cytoid bodies.
> 2. Microaneurysms.
> 3. Roth spots.
> 4. Macular degeneration.
> 5. Nystagmus.

A student responds to item 25 by saying:

> So I'm reading the top. So it says a helpful eye finding in lupus. So actually what I'm doing is, I'm remembering a little mnemonic I learned about findings in lupus and I'm trying to remember if in that mnemonic there is anything to do with eye findings but there isn't. (Probe by interviewer -- So what is the mnemonic?) It's doctor, soap, brain. (O.K.). In the little mnemonic the only eye finding I can remember with lupus is photosensitivity. So what I'll do is I'll go back and look at the answers and see whether they make sense or whether it's something else or not. So I'm looking at number 1; I'm not sure what those are. Number 2, I've read and I'm just thinking about whether that would be possible or not and I think it would be possible. Number 3 ... I've read 3, 4, and 5. I think it's number 2.

In this item the student read the stem, generated an incorrect answer -- *photosensitivity*, and then searched the alternatives for a match. In terms of item-related and successful problem solving activities, these behaviors were coded as A3 and C3. The generated answer was replaced by alternative 2 -- *microaneurysms*. Alternative 1, the keyed answer, was discarded using move B3 -- *Reads an alternative and states "Not sure" or "I don't know"* and the student's interaction with alternatives 3, 4, and 5 was coded as B1 -- *Reads an alternative and makes no comment.*

To show the use of activity B2, item 26 will be used.

> Which of the following glomerulopathies with renal insufficiency is most likely to benefit from corticosteroids or azathioprine?
>
> 1. Goodpasture's syndrome.
>
> * 2. Lupus nephritis.
>
> 3. Pre-eclampsia.
>
> 4. Amyloidosis.
>
> 5. Scleroderma.

In the following think aloud protocol, the student rejected most of the alternatives using activity B2 -- *Reads an alternative and states "No"* when dealing with the alternatives.

> O.K., which of the following glomerulopathies with renal insufficiency is most likely to benefit from corticosteroids or azathioprine? O.K., Goodpasture's syndrome, I'll say no. I can't remember but Goodpasture's doesn't seem right. Lupus nephritis, yea, that's possible but I'm not sure. Pre-eclampsia, the answer is no, amyloidosis, the answer is no, scleroderma, I don't even know if you have a glomerulopathy with scleroderma. I'd go with lupus.

In each of the above protocols, three of the four alternatives were rejected using either activity B1 or B2. However, for any single item, students may use a variety of reasons for not considering an alternative as a plausible answer. The following is a student's interaction with item 25.

> Diagnostically helpful ophthalmic finding in lupus. Roth spots are found in endocarditis; cytoid bodies I haven't heard of. Microaneurysms, macular degeneration, SLE ... I don't know. I'll say microaneurysms, it's just a guess.

This student did not know cytoid bodies (Move B3), alternative 3 is associated with endocarditis (Move B16), alternative 4, macular degeneration, was read but no comment was offered (Move B1), and the last alternative, nystagmus, was ignored (Move B6).

Although frequency counts were not recorded for each alternative, it may be that activities B1 and B2 are used for alternatives with low associative strengths. It is possible that some of the alternatives do not represent the misconceptions held by students with respect to item content or that the alternatives are not homogeneous. As an example, consider the following protocol of a student's response to item 26 and specifically, the student's treatment of the third alternative -- Pre-eclampsia.

> Glomerulopathies so I'm thinking nephrology with renal insufficiencies most likely to benefit from corticosteroids or antimetabolite, well actually that's not antimetabolite, it's DNA inhibitor. So glomerulopathies, I'm thinking glomerulonephritis and nephrotic syndrome, so you can have basement membrane disease, you can have a deposit disease or you can have vasculitic process of the vessels, so the steroids are to

> reduce immune response; they both hit the immune response, they are proliferative of the lymphocytes. So Goodpasture's, that's a basement membrane disease, corticosteroids will help. Lupus, you can use any of these in an acute situation. Scleroderma, well, pre-eclampsia, that's not actually an autoimmune problem, so that's not it. Amyloid, if it's primary or secondary, they don't say. I assume they mean primary, it doesn't help. Scleroderma, it doesn't help, progressive sclerosis. So I would be thinking of the immune phenomenon with Goodpasture's syndrome, blasting them hard with immunosuppressives, works well. With lupus, it's usually a chronic thing and it's going to recur so the person that would most likely benefit would be a person with Goodpasture's ... number one.

In the set of five alternatives, pre-eclampsia is the only response which is not an autoimmune problem and therefore can be eliminated from consideration.

The remaining moves B9, B15, and B16 range in use from 143 to 200 times across the alternatives for the seven items. In contrast to strategies B1 and B2 which were used with the majority of the items, the use of activities B9, B15, and B16 tended to be item-specific. For example, of the 200 times that move B9 -- *Eliminates an alternative using oppositions* was used, it appeared 140 times for the elimination of alternatives in item 8. This item reads as follows.

> Which of the following drugs is generally NOT useful in improving the long term outlook in patients with rheumatoid arthritis?
>
> 1. Penicillamine.
> 2. Gold.
> * 3. Prednisone.
> 4. Chloroquine.
> 5. Methotrexate.

In dealing with drug therapy for a patient with rheumatoid arthritis, students appeared to focus on the phrases "not useful" and "long term outlook". It is inferred that the latter phrase gave rise to the generation of two broad categories of drugs -- those that are disease modifying and those that are symptom relieving. With this framework, students then proceeded to the alternatives. The following student's protocol shows some of these features.

> Which of the following drugs is generally not useful in improving the long term outlook in patients with rheumatoid arthritis? So, you would be looking for disease modifying agents. Penicillamine is a disease modifying agent - - slows the progression of rheumatoid arthritis, gold also slows the progression of rheumatoid arthritis. Prednisone is acute, can be used systemically if these other measures don't control the symptoms and inflammation of rheumatoid arthritis, so that's possible. Chloroquine is also used to control the symptoms of rheumatoid arthritis and it is a disease modifying agent. Methotrexate, they also tend to use. I think Prednisone just relieves the symptoms but it does not actually slow the progression of rheumatoid arthritis, so I think, three, would not improve the long term outlook of patients with RA.

Because the item called for a drug which was not useful, a drug which provided only symptomatic relief would be the correct answer and so the phrase "not useful" gave rise to opposites such as "remitting/nonremitting", "disease modifying/symptom relieving", "useful/useless", "long term/short term", "acute/chronic", "used/not used", "last line/other line", and "plausible/not plausible". These were then used in ruling out and ruling in alternatives.

Oppositions were also used in item 29 (Generalized pruritis is most commonly due to:). Students produced bipolar pairs such as "common/rare" and "would cause/would not cause" to classify a list of conditions and their association with itchiness. In addition to the use of opposites for deleting alternatives in this item, students also used crude likelihoods to rank the causes of generalized pruritis (For example, "Polycythemia, then again it's depending on what is causing the polycythemia but it is unlikely to be the most common cause.").

The majority of the students used activity B15 -- *Eliminates an alternative because it does not match the data provided in the stem*, to discard the alternatives in item 10. Item 10 reads as follows:

> Which of the following series of heart sounds is in the correct sequence? (O.S. = Opening Snap)
>
> 1. 1st, 2nd, 3rd, 4th, O.S.
> 2. 1st, A2P2, 3rd, 4th, O.S.
> 3. 3rd, 1st, A2P2, O.S., 4th.
>
> * 4. 4th, 1st, A2P2, O.S., 3rd.
>
> 5. 4th, 1st, P2A2, O.S., 3rd.

This is an item that presented difficulty in classifying the student strategies. The majority of the students read the entire item and after determining that all the alternatives contained sequences of the first, second, third, fourth, and opening snap heart sounds, generated partial sequences or in a sense partial answers. The following student's protocol illustrates the use of partial sequences to help rule in and rule out alternatives.

> Which of the following series of heart sounds is in the correct sequence where O.S. is the opening snap? If I remember correctly, the opening snap is the sound of the mitral valve as it opens during diastole, about early diastole, so it would occur before an S4 and it would occur also before an S3. So that immediately rules out number 1 because I know that my opening snap must be before the fourth heart sound. Again it rules out number 2 because my opening snap must be before the fourth heart sound. Number 3 has the third heart sound before the first heart sound, so that's not possible. So it's down between number 4 and 5 and again that's showing the correct sequence of an S4 and S1 and then an opening snap and an S3 and now I have to choose which of the S2's comes first, the pulmonic or the aortic. And I know that it's the aortic sound that comes first and then the pulmonic, so my answer is number 4.

In the example, the student developed several rules or partial sequences for the heart sounds. The first of these was that the opening snap should appear before the third and fourth heart sounds. Other sequences that students developed include "Opening snap always comes after the second sound", "Opening snap comes after the second and before the third sound", and "The sequence always begins with the fourth heart sound". Using the partial sequence as a criterion, students eliminated those sequences which failed to meet the criterion. An argument could be made for having these behaviors categorized as B14 -- *Eliminates an alternative because it does not match the generated answer or the answer space.*

Move 16 -- *Eliminates an alternative on the basis of association* was used most frequently in items 2 and 25. Item 2 deals with the most likely physical finding to be noted in a patient during an attack of angina on effort and the following student's protocol to this item shows how association between concepts and alternatives is used to eliminate alternatives.

> The most likely ... (reads silently) ... on effort. First you have to know what angina is. They're asking for angina, so you, my understanding of angina is heart pain brought on due to ischemia of the heart. They ask for the most likely finding. Number one, fourth heart sound, two - decreased second sound, three - mid-systolic murmur at the apex, four - transient aortic ejection click and 5 - third heart sound. Once again, just eliminating ones I feel are probably wrong. Number three - mid-systolic murmur at the apex, that's usually caused by, would be due to probably an aortic stenosis, so I don't think that would be right. A decreased second sound would be due to either the pulmonary valve or the aortic valve, decreased intensity to me would be something like insufficiency, Fourth heart sound is a possibility. A transient aortic click - I doubt it - I don't think that has anything to do with angina. That has more to do with aortic stenosis, more like number three. Third heart sound those are

> usually caused more with atrial fibrillation or flutter or aortic stenosis. Fourth heart sound - choice number one. I'm trying to think what a fourth heart sound is due to. That's usually due to blood flow into an uncompliant ventricle or a piece of mitral that isn't moving. So I think probably number one would be the best answer.

In this protocol the student searched through the alternatives for a physical sign of angina on effort. Each alternative was rejected because the physical sign is not a characteristic of angina. As stated before, each of the alternatives served as a conditional hypothesis. Using the third alternative, mid-systolic murmur at the cardiac apex, as an example, the inferred reasoning of the student might take the form of "if mid-systolic murmur, then aortic stenosis but this is a question on angina on effort so mid-systolic murmur is not right". In such a situation, the student would be using backward reasoning. If the alternatives were treated as conditional hypotheses, then the student worked backward from the unknown (physical signs) to the known (angina on effort).

Association was also used to eliminate alternatives in item 25 which requested a diagnostically helpful finding in lupus erythematosus. In the following protocol, the alternatives, "Roth spots" and "Microaneurysms" are discarded because they are associated with conditions other than lupus erythematosus.

> A diagnostically helpful ophthalmic finding in lupus is. I'm going to read the answers and see what they say. Cytoid bodies, I don't know what those are. Microaneurysms, I'm just going to look at the rest. Roth spots, no because those are more bacterial endocarditis

> and not in lupus. Macular degeneration, I think that's possible. Nystagmus, I don't think so, doesn't seem typical of lupus. It's between number 2, microaneurysms, and number 4, macular degeneration. Microaneurysms seems typical of things like hypertension and stuff. I'm really not sure but I can pretty well rule out ... I don't know what number 1 is so I'm not going to pick it and number 3 I can rule out for sure and number 5 I can rule out just about for sure. Actually, I can't decide between 2 and 4. I'm going to choose 4.

Alternative 1 -- Cytoid bodies was eliminated using move B3 -- *Reads an alternative and states "Not sure" or "I don't know"* while alternative 5 -- Nystagmus was rejected using move B15 -- *Eliminates an alternative because it does not match the data provided in the stem.*

Other activities such as B3 -- *Reads an alternative and states "Not sure" or "I don't know"*, B12 -- *Eliminates an alternative on the basis of priority*, and B13 -- *Eliminates an alternative in terms of likelihood* were used less frequently. The sums for these four activities range from 60 to 91. Activity B12 was used predominantly in item 30 which dealt with a patient with irritable bowel syndrome. Treatment and management plans considered to be inappropriate or not a priority were rejected.

The remaining activities -- B4, B5, B6, B7, B8, B10, B11, and B14 -- were used less frequently and appeared with sums ranging from 1 to 33.

The last section dealt with the activities students used when dealing with the alternatives. Students provided a variety of reasons for eliminating options. Depending upon the content, wording, and task of the items, different reasons or a single reason were used for dispensing with the alternatives in any single item.

For the items contained in Group One, the generation of answers from processing the information contained in the stem only was not a dominant activity. A review of the tabulations for the number of successful problem solving activities (Appendix VIII) shows that neither do students advance hypotheses during the processing of the stems. Forward reasoning, as described by moves C3 and C4 in the present study, did not appear to be a strong characteristic of student-item interaction. Students did not engage in an intermediate activity such as, for example, generating a diagnosis for an item which requests management of a patient, because the item content and task do not necessarily call for such an intermediate activity. On the other hand, successful problem solving behaviors such as restating the stem, focussing on key features and providing a definition, description, or an explanation in terms of the underlying etiology and pathophysiology were used by some students for some of the items. For the majority of the items in this group, answers were selected after a careful consideration of the alternatives. Each alternative served as a provisional hypothesis and it would appear that the student matched, in turn, the features of each alternative with the information presented in the stem and his/her knowledge base until a match was found. In this type of interaction, it is assumed that the student was working from a hypothesis back to the data and therefore the selection of an answer was achieved using backward reasoning.

It would appear that the activities a student evokes in responding to an item results from some complex relationship between the semantic features of the item and the student's knowledge base. It is not necessarily the item's psychometric difficulty that gives rise to the activities a student will summon in answering an item. In fact, psychometric difficulty is perhaps better thought of as an outcome of the engagement and not a property of the item.

Several stem characteristics may account for the observation that students do not generate answers or advance hypotheses for the items in Group One. To review these, the stems of the seven items are reproduced below.

> Item 2
> The most likely physical finding to be noted in a patient during an attack of angina on effort is a:
>
> Item 8
> Which of the following drugs is generally NOT useful in improving the long term outlook in patients with rheumatoid arthritis?
>
> Item 10
> Which of the following series of heart sounds is in the correct sequence? (O.S. = Opening Snap)
>
> Item 25
> A diagnostically helpful ophthalmic finding in lupus erythematosus is:
>
> Item 26
> Which of the following glomerulopathies with renal insufficiency is most likely to benefit from corticosteroids or azathioprine?

Item 29
Generalized pruritus is most commonly due to:

Item 30
The irritable bowel syndrome in adults ("irritable colon") is a diagnosis of exclusion. However, when this diagnosis is finally made you should:

One of the characteristics of a properly constructed item is that the stem could be administered as a constructed-response item and that the best or correct answer could be provided in the absence of the alternatives. Although most of these stems present a task for the examinee, the generality of the task may preclude the generation of an answer. For item 2, the student may be thinking that there is a large number of physical signs of angina; there are signs which may be observed from the general appearance of a patient and other signs which may be obtained on palpation and auscultation. Likewise, in item 30, there may be many things one should do after diagnosing a patient with irritable bowel syndrome. Because the stem fails to present a problem, the answer space may be large and the student may be unable to reduce the space to a more manageable size. Making the stems more specific, for example, in item 2 requesting for a physical finding on auscultation and in item 8, asking for medical management of a patient with irritable bowel, may lead to the generation of appropriate physical findings and medical interventions respectively.

A similar argument of an unmanageable answer space can be advanced for items 10, 26, and 29. In item 10, students do not know whether the sequences of heart sounds in the alternatives contain all or only a subset of the heart sounds. The list of glomerulopathies with renal insufficiency (Item 26) may be more than four or five and in that list there may be several which could be managed by corticosteroids or azathioprine. Generalized pruritis (Item 29) can be caused by many things.

Items 2, 8, 26, and 29 contain relative phrases such as "most likely", "most commonly" or "generally". As far back as 1950, Bloom and Broder stated that students experience difficulties in determining an answer for items which contain relative terms. The use of phrases like "most commonly" may lead a student to focus on the meaning of the relative phrase. This was the case for one student who in responding to item 29 said "This is kind of silly question because you can always think of a billion reasons when it says most commonly ...".

Item 8 contains a negative. The task for the examinee is to select a drug which is not useful as a long term pharmacological intervention for rheumatoid arthritis. The number of drugs which would not be used for a patient with this condition is indeed large. There are at least fifteen different nonsteroidal anti-inflammatory drugs which would be administered in the early stages but would not be used for long term effects (The Medical Letter on Drugs and Therapeutics, 1991). In addition to these fifteen drugs, there are other unrelated drugs (for example, drugs used to control high blood pressure) which would not be used in managing a patient with rheumatoid arthritis. From the student's perspective, the more efficient way of responding to this item may be to work through the alternatives looking for a drug which is not useful rather than generating an answer of a not useful drug and then searching for a match. Once again, Bloom and Broder (1950) state that the negatively worded item requires the student to make a readjustment of his or her mental set. Because the majority of items used in this study as well as the majority of items appearing on any exam are positive, students are oriented toward the selection of true or positive statements.

Finally, students may experience difficulty in generating an answer because the item is not appropriate for the students' level of training. As an example of an item's relevance, consider item 25 which requests an ophthalmic finding for diagnosing lupus. According to subject matter experts the item is more appropriate for the postgraduate level of medical training. Of the forty students responding to this item, twenty-seven gave a response of "do not know" to the alternative keyed as the best answer.

In summary, factors such as general and unfocussed stems, the use of negation and relative phrases, item content, and intended behavior to be tested by the item may hinder the students in generating answers and advancing hypotheses.

Global Patterns and Flowchart for Group One Items

Global patterns used by the students to respond to the items in Group One are shown in the form of a flowchart presented in Figure 2. The patterns are developed primarily from the item-related activities which deal with generating an answer and the two successful problem solving moves associated with producing an answer and activating a hypothesis. In this set of items, there is a total of 279 (40 students X 7 items - 1 student-item engagement) instances for the generation of answers upon reading the stem only. In general, students begin by reading the stem. The right hand side of the flowchart (double-lined arrow and double lines) reflects

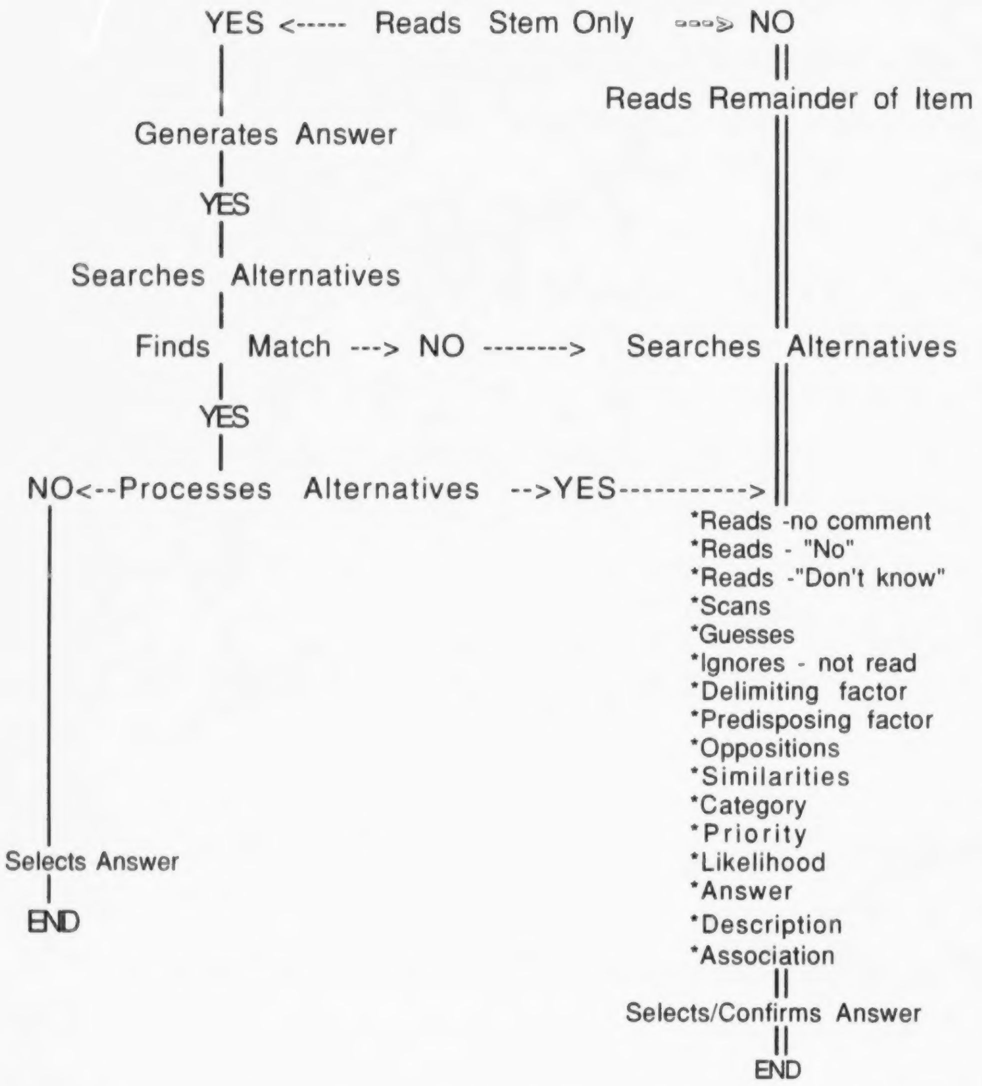
YES <----- Reads Stem Only ⇒ NO
Reads Remainder of Item
Generates Answer
YES
Searches Alternatives
Finds Match ---> NO --------> Searches Alternatives
YES
NO<--Processes Alternatives -->YES----------->
*Reads -no comment
*Reads - "No"
*Reads -"Don't know"
*Scans
*Guesses
*Ignores - not read
*Delimiting factor
*Predisposing factor
*Oppositions
*Similarities
*Category
*Priority
*Likelihood
*Answer
*Description
*Association
Selects/Confirms Answer
END
Selects Answer
END

FIGURE 2: Flow chart for "few or no answers generated; no hypotheses activated" items

the route taken by the majority of students. Of the 279 occasions for generating answers which match in form one of the item alternatives, there were 262 instances where no answers were produced. For this set of items, students generally do not generate an answer but instead read the entire item and select an answer by processing the alternatives. It would appear that the selection of an answer is triggered by the alternatives.

The left hand side of the flowchart represents the path taken by students who generated an answer matching in form one of the item alternatives. The number of students who followed the left hand side of the flow chart to answer the items in Group One is small. There were 17 student-item engagements in which an answer was produced upon reading the stem only and prior to the reading of the item alternatives. This is followed by a search of the alternatives for the answer. The search either confirms the produced answer or the alternatives trigger a more suitable answer. In this search mode, distractors are rejected with reason. If, in generating answers, students complement these moves with other activities which enable them to internalize the information and develop problem or answer spaces from the information contained in the stem, then forward reasoning is demonstrated. If the rejection of options is based on the interpretation of the information contained in the stem, then backward reasoning is being used to confirm the answer.

In reading the items, these students may, for example, interpret the information by providing the etiology and pathophysiology. Their behavior would be similar and in some cases identical to the students who produced answers. The difference is, that students following the right hand side of the flow chart do not produce answers on reading the stem only. In dealing with the alternatives, students use a variety of reasons for dispensing with those that are incorrect.

Responding to the items encompasses activities which are not captured by the flowchart. For example, expressing the stem in one's own words and interpreting the meaning of the information presented in the stem and alternatives are activities which students displayed but do not appear in the flowchart.

Strategies for Group One Items

In summary, the results for the Group One items show that students use different approaches in their encounters with multiple choice items. For this set of items, the examinee-initiated actions or moves associated with the item-related and disposition of alternatives identified in Chapter III are combined to produce one of three general strategies. In the first strategy, students read the stem, go on to the alternatives in search of an answer, and consider the plausibility of the alternatives as possible answers. The second strategy consists of students reading the stem and the alternatives as a unit, proceeding then to a more detailed search and processing of the alternatives for an answer. In the third approach, students read the stem, generate an answer, and then go on to the alternatives in search of the generated answer or a better answer. Incorporated in each of these strategies is a component related to the processing of the alternatives. For the items in Group One, there are 1,114 student-distractor engagements. For about two-thirds of these engagements, reasons reflecting knowledge and thinking were provided for rejecting incorrect alternatives. For the remaining one-third, no real reasons were given. That is, these were engagements that were coded as move B1, B2, B3, or B6.

Coupled with each of these strategies is a component made up of successful problem solving moves. Upon reading the stem or the entire item students may restate the item in their own words, provide a more in depth description by referring to etiology and pathophysiology, develop categories of diseases, or advance an answer. Of course, the extent to which students restate or redescribe the information presented in the item depends on the cognitive demands made by the item. If the students produce these descriptions by building on the information presented in the stem or entire item, an argument can be made that forward reasoning is being used. However, this is only a partial representation of how the items are answered. As stated earlier, incorrect options are usually rejected with reason. Each alternative serves as a hypothesis and its suitability as an answer is determined by referring either to the information presented in the stem or to the student's internalization of the problem. In such instances, the discarding of alternatives is done through backward reasoning.

For situations where forward reasoning is exhibited, the student is behaving like an expert but the format of the multiple choice item places some demands on the student. Alternatives are checked to determine if something has been overlooked and so the plausibility of each alternative is reviewed. This check loop is backward oriented.

The next section of Chapter IV deals with the second set of items.

Group Two Items

The second set consists of items 3, 7, 12, 13, 17, 18, and 19. For these items, students generated few or no answers but activated hypotheses. To review, in this study the activities of generating answers (Moves A3 -- *Reads stem, generates answer, searches for answer, alternative triggers different answer, discards initial answer and eliminates other alternatives,* A4 -- *Reads stem, generates answer, searches, selects generated answer, scans or ignores remaining alternatives,* A5 -- *Reads stem, generates answer or answer space, searches, selects, eliminates alternatives, provides rationale for most of the eliminated alternatives*, and C3 -- *Generates answer or answer space using the information presented in the stem*) and activating hypotheses (Move C4 -- *Activates hypothesis using the information presented in the stem*) on the basis of the information presented in the stem are taken as indicators of forward reasoning. The distinction between the two activities is somewhat fuzzy and dictated by the task set for the examinee in the stem. To clarify this, consider item 12 in this set.

> A 75-year-old man gives a long history of intermittent diarrhea, some constipation, and lower abdominal pain. Recently the pain has become more severe and has been associated with urinary frequency and burning. On

> physical examination the temperature is 39°C and a tender mass is palpable in the left lower quadrant of the abdomen.
>
> A complication likely to develop in this patient is:
>
> * 1. Vesico-colic fistula.
> 2. Massive hemorrhage.
> 3. Subphrenic abscess.
> 4. Jaundice.
> 5. Urinary obstruction.

This item begins with a description of an elderly man who presents with a number of signs. However, rather than asking for a diagnosis, the task for the student is to select a complication. Here is how one student processed the information contained in the item's stem.

> A 75-year-old man gives a long history of intermittent diarrhea, O.K., it's been going on for a long time, some constipation and lower abdominal pain, kind of vague symptoms. Recently, the pain has become more severe and has been associated with urinary frequency and burning, so I'm thinking more of a urinary tract infection. Dull pain related to the diarrhea and constipation. On physical examination, the temperature is 39, he's febrile which is nice to know, some type of infection and a tender mass is palpable in the left lower quadrant of the abdomen. Complication likely to develop in this patient is: nothing comes to mind on what his diagnosis is and they are assuming you have a diagnosis. So he has got intermittent diarrhea, constipation and lower abdominal pain. So he might be getting Crohn's or ulcerative colitis. Why has the pain become more severe with urinary frequency and burning? Unless this guy has got Crohn's, recently diagnosed Crohn's and he has got a fistula into

> his bladder causing urinary tract infection. He is febrile, tender mass in the left lower quadrant. That would go along with inflammatory bowel disease. Complication likely to develop in this patient ... .

In this think aloud protocol, the student offered a diagnosis (Move C4) of urinary tract infection on the basis of a set of signs and symptoms consisting of lower abdominal pain, severity of pain, urinary frequency, and burning sensation on urination. The temperature of 39 supports the infection. The findings of intermittent diarrhea, constipation and lower abdominal pain support a bowel condition such as Crohn's disease or ulcerative colitis (Move C4). The student connected the broad diagnosis of inflammatory bowel disease with the diagnosis of urinary tract infection and then generated an answer, fistula, (Move C3) as a possible complication.

Except for item 17, each of the items in this group begins with a scenario consisting of patient information such as age, gender, presenting complaints, signs, symptoms, and results of laboratory tests. Usually, more than one body system is involved. The main focus of each item is not on diagnosis but on other aspects such as x-ray findings, complications, next course of action, and management. Activating a hypothesis or putting forward a diagnosis represents an intermediate step in resolving the patient's problem.

A review of the frequency of activity use (Appendix VIII) for the item-related activities shows that move A2 -- *Reads stem, searches for answer among alternatives, answer triggered by alternative, eliminates remaining alternatives* was the dominant activity used by the students in reaching an answer. Activities associated with the production of an answer prior to reading the alternatives (Moves A3, A4, and A5) and move A1 -- *Reads stem and all alternatives* appeared infrequently. With respect to the item-related activities, the students' approach to this set of items did not differ very much from their approach to the items in Group One. However, there are differences concerning the frequency with which successful problem solving activities were used. For the Group Two items, students used moves C1 -- *Restates information presented in the stem in own words* and C2 -- *Focusses on key features and defines or redescribes in a different manner* more frequently and consistently across the seven items. Because of the nature of the items, activities C1 and C2 give rise to move C4 -- *Activates hypothesis using the information presented in the stem.* In such cases, a student's interpretation of the information presented in the patient scenario was captured by a diagnosis.

As with Group One items, students used a variety of reasons for rejecting alternatives. This is reflected in Table 7. Of the sixteen activities related to the rejection of options, moves B1 -- *Reads an alternative and makes no comment*, B2 -- *Reads an alternative and states "No"*, B3 -- *Reads an alternative and states "Not sure" or "I don't know"*, B12 -- *Eliminates an alternative on the basis of priority*, B13 -- *Eliminates an alternative in terms of likelihood*, B15 -- *Eliminates an alternative because it does not match the data provided in the stem*, and B16 -- *Eliminates an alternative on the basis of association* were used most frequently.

Table 7

Frequency of Move Use for Alternatives of Group Two Items

		Items							
	MOVE	3	7	12	13	17	18	19	Sum
1.	No comment	21	12	11	17	14	17	21	113
2.	No	7	7	20	14	1	15	13	77
3.	Do not know	8	-	5	1	13	31	9	67
4.	Scans	-	-	-	4	-	8	4	16
5.	Guesses	-	-	-	-	-	-	-	-
6.	Ignores	1	1	5	3	-	1	15	26
7.	Delimiting	-	-	-	3	-	-	1	4
8.	Predisposing	-	-	-	-	-	-	1	1
9.	Oppositions	5	2	-	-	-	1	-	8
10.	Similarities	-	2	-	6	8	-	-	16
11.	Category	1	5	-	3	-	-	-	9
12.	Priority	-	84	-	76	49	42	9	260
13.	Likelihood	9	5	39	6	1	5	17	82
14.	Answer	-	-	-	2	-	-	-	2
15.	Description	47	6	22	6	-	4	21	106
16.	Association	61	36	58	19	74	36	49	333
	TOTAL	160	160	160	160	160	160	160	1120

The activities used most frequently for rejecting alternatives for Group Two items are similar to those used by the students for discarding options for Group One items.

To show the use of the various item-related, disposition of alternatives, and successful problem solving activities for Group Two items, examples of student think aloud protocols follow. The first of these is a student's interaction with item 3. The item reads as follows.

> A 65-year-old male who has a history of heavy cigarette consumption for many years presents with pulmonary emphysema and intercurrent acute bronchitis. In addition to the signs of chronic lung disease, he has massive peripheral edema, enlarged pulsating liver, elevated jugular venous pressure, and large V waves in the neck.
> His x-ray may be expected to show:
> 1. Acute congestion of the pulmonary veins.
> 2. Enlarged left ventricle with increased aeration of the lungs.
> 3. Straightening of the right border of the heart by right ventricular enlargement.
> 4. Bi-ventricular enlargement and normal pulmonary vessels with flattening of the diaphragm.
>
> *5. "Pruning" of smaller branches of the pulmonary artery and flattening of the diaphragm.

In the following protocol, the student presented several answers based on a diagnosis formulated from the data presented in the stem.

> A 65-year-old male, history of cigarette consumption for many years presents with pulmonary emphysema and intercurrent acute bronchitis. So we know he has two things here plus the heavy cigarettes. In addition to the signs of chronic lung disease, he has massive peripheral edema, O.K., a sign of heart disease could be enlarged pulsating liver, elevated jugular venous pressure, and large V waves in the neck. All these things point me to congestive heart failure of the right side of the heart. His x-ray may be expected to show, well, right sided heart failure is what I see so I might look for something like a big heart, enlarged heart shadow, so let's see if we have anything like that here.

Using the information presented in the stem, the student associated massive peripheral edema with heart disease. The additional signs of enlarged pulsating liver, elevated jugular venous pressure, and large V waves helped the student refine the diagnosis to one of congestive heart failure of the right side of the heart. These activities would indicate that the student was using move C4 -- *Activates hypothesis using the information presented in the stem.* Because the student expressed the clinical presentation in a form that has personal meaning, the successful problem solving activities C1 -- *Restates information presented in the item in own words* and C2 -- *Focusses on key features and defines or redescribes in a different manner* were also being exhibited.

A diagnosis of right sided heart failure led the student to generate x-ray findings of big heart and enlarged heart shadow (Move C3 -- *Generates answer or answer space using the information presented in the stem*). Several aspects of this protocol bear some of the features of expert and successful problem solving. The student generated a diagnosis. However, the list of possible diagnoses and the clarification of the differential diagnosis were incomplete because the lung conditions and the heart problems were not integrated. The information presented in the stem was expressed by the student in a way that had personal meaning and understanding. The student supplemented the problem space, that is, the diagnosis of right sided heart failure, with responses that related directly to the requested task of identifying x-ray findings. By generating a short list of x-ray findings, the student developed a set of answers that were used in a search of the alternatives for a possible match. The combination of moves related to activating a diagnosis from signs and then generating answers on the basis of the diagnosis suggests that there are some elements of forward reasoning in this student's think aloud protocol.

Because students engaged in an intermediate activity of formulating diagnoses or hypotheses for this set of items, it is also possible to consider the think aloud protocols in terms of the diagnostic problem solving approaches proposed by Ramsden, Whelan and Cooper (1989). In the case of the above protocol, the student used short associative links to relate patient clinical features to a diagnosis of congestive heart failure. Using the classification offered by Ramsden *et al*, at this stage the student would be using an ordering approach to reaching a diagnosis. The approach is not a structured one because the student failed to integrate the heart and lung findings. The student then turned to the alternatives.

> Acute congestion of the pulmonary veins, more likely to be due to left heart failure, enlarged left ventricle with increase aeration of the lungs is the same thing. Straightening of the right heart border of the heart by right ventricular enlargement looks right. Bi-ventricular enlargement and normal pulmonary vessels with flattening of the diaphragm, possible because of the pulmonary emphysema would definitely give you flattening of the diaphragm. Pruning of the smaller branches of the pulmonary artery and flattening of the diaphragm, O.K., so we ruled out the first two, now it's the last three. It doesn't say here that he is short of breath so I would like to think ... well, he actually ... just

> because he has bronchitis and pulmonary emphysema, he could have bi-ventricular enlargement, for this fellow I would probably have to say four.

Move B16 -- *Eliminates an alternative on the basis of association*, was used to reject the first two options because the x-ray findings are associated with left heart failure. The third and fourth options are, according to the student, possible answers. A rationale for the suitability of the third alternative was not provided, however, the fourth option was justified because one of the x-ray findings, flattening of the diaphragm, is associated with pulmonary emphysema. Option five was read but no overt action was taken (Move B1 -- *Reads an alternative and makes no comment*). The fourth alternative was selected as an answer. Because the student generated possible x-ray findings (answers) such as "big heart", and "enlarged heart shadow" and provided reasons for most of the rejected options, move A3 -- *Reads stem, generates answer, searches for answer, alternative triggers different answer, discards initial answer and eliminates other alternatives* characterizes this student's overall approach to the item.

If each of the options is considered as a provisional hypothesis, then with respect to this student's selection of an answer and the rejection of options, several additional features emerge. First, backward reasoning was evidenced in the rejection of the first two options and the selection of the answer. That is, the student searched for an alternative in which the x-ray findings are characteristic of right sided heart failure but the findings presented in the first two options are associated with left sided failure and are, therefore, rejected. Rejection of the first two options was by exclusion. Second, partial information was used to arrive at an answer. In the first encounter with option four, the student took only one of the x-ray findings (flattening of the diaphragm) and related it to one of the conditions (pulmonary emphysema) presented in the stem. At the stage where the student confirmed the fourth option as an answer, the x-ray finding, bi-ventricular enlargement, was related to the two conditions of bronchitis and pulmonary emphysema. Not only did the student use partial data to confirm the choice of answer but the information that was used differed between the first encounter with option four and its selection as an answer. In terms of the diagnostic strategies identified by Ramsden *et al* (1989), the selection of an answer by this student would be labelled as a pattern matching strategy. It should also be noted that in reading the stem, the student did not provide overt evidence for understanding either the emphysema or the bronchitis. No meaning or interpretation of these two conditions was offered. Neither was there any attempt to relate these two conditions to the physical findings associated with the heart failure and the diagnosis right sided congestive heart failure. Once again, within the framework provided by Ramsden *et al*, this student's approach to the item would be labelled as ordering. A third feature of this protocol is that forward reasoning does not necessarily lead to the correct answer, although it could be argued that the reasoning was incomplete because the student did not tie the lung and heart findings.

Even when the pattern of reasoning appears to be very similar for several students, the keyed answer was not necessarily selected. Consider the protocols of two students as they dealt with the information presented in item 3.

> A 65-year-old male, so he's old. He has a risk factor for coronary heart disease, who has a history of cigarette consumption. He has another risk factor, for many years presents with pulmonary emphysema, O.K., so he has COPD and intercurrent acute bronchitis. So he has the spectrum of COPD. In addition to the signs of chronic lung disease, he has massive ... oh, O.K., this is cor pulmonale, he has massive peripheral edema, enlarged pulsating liver, elevated JVP, and large V waves in the neck. O.K., so this is cor pulmonale, right ventricular failure, secondary to the chronic bronchitis. His x-ray may be expected to show -- acute congestion of the pulmonary veins, no, this is right heart failure, not left heart failure. Enlarged left ventricle with increased aeration of the lungs, I doubt that; this is right heart failure. Straightening of the right border of the heart by right ventricular enlargement, that's very possible. Bi-ventricular enlargement, O.K., no, it wouldn't be bi-ventricular because this is clearly right heart failure. Chronic bronchitis causes cor pulmonale, right heart failure, so peripheral edema and right heart failure, pulsating liver, right heart failure, JVP is right heart failure, V waves in the neck, right heart failure. So 1 doesn't make sense, 2 doesn't make sense, 4 doesn't make sense. Five, pruning of the smaller branches of the pulmonary artery and flattening of the diaphragm, pruning of the smaller branches of the pulmonary artery, I think 3 is a much better answer. Clearly cor pulmonale and right ventricular enlargement.

This student advanced the diagnosis of cor pulmonale and qualified the heart disease as right ventricular failure brought on by the chronic bronchitis (Move C4 -- *Activates hypothesis using the information presented in the stem*). Three of the options were rejected because the x-ray findings are not associated with right heart failure (Move B16). The fifth alternative was read and rejected because option 3 is a better or a more likely answer (Move B13) and had been considered as a possible answer earlier. An answer to the item was triggered by the alternatives (Move A2).

A similar approach was used by the second student. Here is how this student handled the stem.

> Reading the stem -- 65-year-old male, heavy cigarettes, presents with emphysema and bronchitis. In addition to his chronic lung disease, massive peripheral edema, enlarged, pulsatile, pulsating liver, elevated JVP and large V waves. These are all, not all, pardon me, the signs in the last sentence immediately bring to mind right sided heart failure and the first part of the stem sets up a COPD patient with right sided heart failure. That makes me think right away of cor pulmonale and I haven't read the question yet. So, the question is his x-ray may be expected to show. So I have a diagnosis of cor pulmonale and they are asking for x-ray findings and in this setting I would go and read the answers and see which answers fit.

Upon reading the stem only, the second student formulated a diagnosis of cor pulmonale (Move C4). Compared to the first student who qualified the heart failure to be right ventricular and secondary to the lung problems, the second individual stated that the heart condition is right sided. Because both students offered a diagnosis of cor pulmonale, it is inferred that the students were able to integrate the patient's clinical features pertaining to the pulmonary and cardiovascular systems. These actions bear resemblance to the structured approach of reaching diagnoses described by Ramsden *et al* (1989). In the strictest sense, whether the students used a stepwise pathophysiological or a diagnostic integration strategy is questionable. According to Ramsden *et al*, these strategies would be evidenced if students used pathophysiology to explain the patient's signs and symptoms. For the purposes of this thesis and indeed from the perspective of typical item performance, such explanations were neither requested nor required of the students.

Having dealt with the information presented in the stem, the student then went on to the alternatives.

> O.K., acute congestion of the pulmonary veins is a finding in left sided heart failure, so that doesn't work. Enlarged left ventricle, again, that's a setting of left sided failure which doesn't fit in here. Straightening of the right heart border by right ventricular enlargement, well, maybe I'll consider that, we will see what else there is. Bi-ventricular enlargement, normal pulmonary vessels and flattening of the diaphragm. Yes, he should have flattening of the diaphragm. There is no reason for his left ventricle to be enlarged given the information we're given here and his pulmonary vasculature should be normal. Number 5, pruning of the smaller branches of the pulmonary artery and flattening of the diaphragm. Those are in fact compatible with the diagnosis of cor pulmonale, the flattening of the diaphragm with COPD

> and the pruning of the pulmonary artery is compatible with emphysema. So I know number 5 is right, so I don't even bother going back to number 3 and I choose number 5.

Both students showed similar patterns of reasoning when dealing with the options. Alternatives 1, 2, and 4 were rejected because of their association with left sided heart failure (Move B16 -- *Eliminates an alternative on the basis of association*). On first pass, option 3 was considered a possible answer. Differences between the students lay with the selection of an answer. The first student offered a very limited rationale for the elimination of the fifth option, the keyed answer. It was rejected because option 3 was a better answer. The second student selected the keyed answer by relating the x-ray findings to the information presented in the stem and the generated diagnosis of cor pulmonale. Reasons were provided for the rejection of the options. In terms of the item related activities listed in Chapter III, this student's overall approach to this item can be characterized by move A2 -- *Reads stem, searches for answer among alternatives, answer is triggered by alternative, eliminates remaining alternatives.*

Of the forty students, thirty-nine generated a diagnosis (right heart failure, right ventricular failure, congestive heart failure, cor pulmonale) for item 3. However, only fourteen students selected the keyed answer. Although students appear to behave like experts by exhibiting elements of forward reasoning, selection of the best answer is not guaranteed. Selecting an incorrect answer for item 3 may be due, in part, to incomplete forward reasoning and a failure to internalize the problem completely by not uniting the diagnosis associated with the lung findings with the findings leading to the diagnosis of heart failure.

The most frequently used activities for rejecting the options in item 3 were moves B15 -- *Eliminates an alternative because it does not match the data provided in the stem* (For example, "Enlarged left ventricle, they haven't given us any information that he would have that.") and B16 -- *Eliminates an alternative on the basis of association* (For example, "Acute congestion of the pulmonary veins, well, if he had left heart failure you would expect congestion of the pulmonary veins.").

Because most of the Group Two items consist of scenarios which present patient clinical features related to several body systems, students are required to integrate this information. The integration represents an intermediate activity and appears in the form of differential diagnoses which should account for all or most of the relevant clinical features. The formation of differential diagnoses is an activity that the majority of students demonstrated. However, as was shown in the think aloud protocols of the two students responding to item 3, evoking a diagnosis did not necessarily lead to the correct answer. Reaching the correct answer may be attributed, in part, to the degree to which a student is able to integrate the clinical features presented in the scenario.

Additional examples of students' think aloud protocols for other Group Two items follow. The first item in this set of protocols is item 7 which reads as follows.

> A 70-year-old man has had a smoker's cough for many years. Two weeks ago his voice became weak and hoarse, and one week ago he began to cough up 50 to 100 g of purulent sputum daily. A chest x-ray reveals a cavity containing an air-fluid level in the left upper lobe. He is afebrile.
>
> Which of the following would you advise?
>
> 1. Surgery with removal of the affected lobe.
> * 2. Bronchoscopy.
> 3. Sputum for acid fast bacilli culture daily for three days.
> 4. Intensive antibiotic therapy.
> 5. Postural drainage four times daily.

In the first example, as the information in the stem is read, the student evokes several hypotheses.

> A 70-year-old man has a smoker's cough for many years so he might have chronic bronchitis. Two weeks ago his voice became weak and hoarse which might signify a laryngeal nerve problem and one week ago he began to cough up 50 to 100 grams of purulent sputum daily. Possibly showing some type of infection and abscess or severe pneumonia. Chest x-ray reveals a cavity containing an air fluid level in the upper lobe. He is afebrile. An air fluid level could signify a pneumatocele, staph aureus infection, could be TB, possible TB infection. Which of the following would you advise?

This student activated at least five differential diagnoses (chronic bronchitis, laryngeal nerve problem, pneumonia, staph aureus infection, and TB). In terms of the successful problem solving activities identified in this thesis, the student exhibited moves C2 -

- Focusses on key features and defines or redescribes in a different manner and C4 -- *Activates hypothesis using the information presented in the stem.* However, it should be noted that the clinical features along with their associated diagnoses were not integrated in such as a manner as to preserve the inherent structure of the problem with which the patient presents. Clinical features were associated with each diagnosis using a pattern matching approach. Since an answer was not generated, this student used the alternatives to trigger an answer. The student then proceeded to the alternatives.

> Which of the following would you advise? Surgery with the removal of the affected lobe. That doesn't fit to be a very good answer -- to remove a whole lobe because of an abscess or an infection. Bronchoscopy may be useful if you if you can't get sputums to diagnose what he actually has. Sputum for acid fast bacilli culture daily for three days, that may be useful depending on the clinical scenario and his contacts. Intensive antibiotic therapy, that would probably be warranted in this case because air fluid level often signifies a staph aureus infection. And postural drainage four times daily may be useful as an adjutant measure but I think the best answer would be number 4, intensive antibiotic therapy, if this is a staph aureus infection.

The first and fifth alternatives were rejected using move B12 -- *Eliminates an alternative on the basis of priority.* Initially, alternatives 2 and 3 were considered as possible answers but were replaced by a more likely one. Therefore, alternatives 2 and 3 were rejected on the basis of move B13 -- *Eliminates an alternative in terms of likelihood.* The fourth option was selected as an answer because it represents the appropriate treatment for the diagnosis, staph aureus infection, which was forwarded by the student. Because the student did not generate any answers, the selection of an answer was triggered by the alternatives. Options considered inappropriate by the student were eliminated with reason. This student's overall approach to this item can be characterized by move A2 -- *Reads stem, searches for answer among alternatives, answer is triggered by alternative, eliminates remaining alternatives.*

The next think aloud protocol for item 7 shows a student taking a different approach from the previous student.

> Again, I'll read this question out loud. A 70-year-old man has a smoker's cough for many years. Two weeks ago, his voice became weak and hoarse and immediately my mind starts to think laryngeal cancer and things like that. One week ago he began to cough up 50 to 100 grams of purulent sputum daily. Chest x-ray reveals a cavity containing an air fluid level in the left upper lobe so now I have changed my sort of thought process thinking more of a lung tumor with recurrent laryngeal nerve involvement. He is afebrile. Which of the following would you advise?

In contrast to the first student who generated five differential diagnoses, this student began by evoking the diagnosis laryngeal cancer. Then on the basis of the additional data on sputum and the chest x-ray, the original diagnosis was refined to one of a lung tumor with recurrent laryngeal nerve involvement (Move C4). This student integrated the clinical features into a single working diagnosis and then went on to the alternatives.

> Which of the following would you advise? Surgery with the removal of the affected lobe, well, I would sort of rule that out right now because you would want to do more investigations before you operate on this man. Bronchoscopy sounds like a good choice. Sputum for acid fast bacilli and culture daily for three days, that may be a good choice but I would still want to rule out carcinomas. That wouldn't be my first choice. Intensive antibiotic therapy, again, I would still want to be ruling out carcinoma. Postural drainage four times a day. So, I think looking at the choices given here, bronchoscopy would be the most reasonable simply because you would try to get a diagnosis of maybe a lung cancer causing the infective process.

Option one was rejected on the basis of priority (Move B12 -- *Eliminates an alternative on the basis of priority*); the fifth option was read but no reason given for its rejection (Move B1 --*Reads an alternative and makes no comment*) while alternatives 3 and 4 were rejected using move B16 -- *Eliminates an alternative on the basis of association.* The student sought an answer that is associated with confirming a diagnosis of lung cancer and options 3 and 4 do not meet this requirement. The second option was selected as an answer because doing a bronchoscopy would enable a ruling in or a ruling out of lung cancer. With respect to the item related activities, this student's overall approach would be characterized by move A2 --

Reads stem, searches for answer among alternatives, answer is triggered by alternative, eliminates remaining alternatives.

One of the aspects that emerges from the protocols presented for items 3 and 7 is the difference in diagnostic strategies. Although students were able to show some of the characteristics of successful problem solving, such as expressing the clinical information in their own words, supporting redescriptions with, for example, pathophysiology, creating an internal problem or answer space, and using forward reasoning, their approaches for selecting an answer differed. With reference to the two think aloud protocols for item 7, the first student used short associative links to relate usually one clinical feature to a diagnosis. This resulted in the activation of many diagnoses. The student failed to integrate all the clinical information, decisions were made using a portion of the data presented, and an incorrect answer was selected and justified because it related to an evoked diagnosis. The evoked diagnosis was associated with one of the patient's clinical features.

The second student's approach to item 7 was more structured. A single diagnosis was initially forwarded and was refined as more patient information was presented. Most of the clinical features were explained in terms of the generated diagnosis. The selection of an answer was based on whether it would help rule in or rule out the lung tumor. By relating the clinical features to the diagnosis and linking the answer to the diagnosis, the student was able to internalize or represent the problem in a way that its inherent structure was maintained. That is, this is a patient who may have a lung tumor with recurrent laryngeal nerve involvement. This is different from the first student who activated at least five diagnoses.

For item 7, the most frequently used activities for rejecting the options were moves B12 (*Eliminates an alternative on the basis of priority*) and B16 (*Eliminates an alternative on the basis of association*). Activity B12 was usually used with options 1 and 5. The following example shows a student using move B12 to eliminate option 1.

> So number one, surgery with removal of the affected lobe. Well, I think you have to do some initial tests before you start removing lobes, so I don't think that would be the first thing.

In the following example, move B16 was used by the student to eliminate option 4, intensive antibiotic therapy. In the beginning part of the protocol, the student discarded a diagnosis of pneumonia because the patient did not have a fever.

> He is afebrile so it's probably not something like pneumonia. ... Intensive antibiotic therapy, no, because I don't think he has a pneumonia.

The student then associated intensive antibiotic therapy as a treatment for pneumonia but since the patient does not have pneumonia, option 4 was not an appropriate action.

The differences between students in diagnostic strategies and approaches to answering the items observed in the think aloud protocols for items 3 and 7 were also evident in other items in this set. The next think aloud protocol is for a student who advanced a hypothesis and generates several answers for item 12. The item reads as follows.

> A 75-year-old man gives a long history of intermittent diarrhea, some constipation, and lower abdominal pain. Recently the pain has become more severe and has been associated with urinary frequency and burning. On physical examination the temperature is 39°C and a tender mass is palpable in the left lower quadrant of the abdomen. A complication likely to develop in this patient is:
>
> * 1. Vesico-colic fistula.
> 2. Massive hemorrhage.
> 3. Subphrenic abscess.
> 4. Jaundice.
> 5. Urinary obstruction.

The student's interaction with the information presented in the stem follows.

> The stem, 75-year old man, long history of intermittent diarrhea, constipation and lower abdominal pain. Pain is getting worse and is associated with urinary frequency and burning. On exam, he is febrile and has a tender mass in his left lower quadrant. Well, you think about what the possible diagnosis is here after reading the stem and I would say at this point, I would have a choice to go on and read the question or to go back and read the stem and try to make a diagnosis. And I'll read the question to see if a diagnosis is required. And the question ... what's the most likely complication? Therefore, we need to know what the diagnosis is before we can answer the question. So, yes, we have to go back and read the stem again and make a diagnosis. You have an older guy, diarrhea,

> constipation, pain. Pain is now more severe, associated with urinary tract symptoms and on physical exam he is febrile and has a tender mass in his left lower quadrant. That sounds to me like someone with diverticular disease, given the physical findings and the epidemiology of the patient. So, therefore, I'll make the diagnosis of diverticular disease and now go look for a likely complication of diverticular disease. When I think of the complications of diverticular disease before I go to look for the answer, I think of things like hemorrhage, obstruction, perforation and then subsequent complications that can develop from those. Maybe I'll just go with what the choices are.

Thus far in the protocol, the student focussed on the key features and expressed the patient information in his or her own words (Move C2 -- *Focusses on key features and defines or redescribes in a different manner*), advanced a diagnosis of diverticular disease (Move C4 -- *Activates hypothesis using the information presented in the stem*), and generated "hemorrhage, obstruction, perforation" as possible answers (Move C3 -- *Generates answer or answer space using the information presented in the stem*). The student then dealt with the alternatives.

> A likely thing to develop, well, visico colic fistula, number 1 is possible but it's not particularly likely unless you are postulating that his urinary frequency and burning is there because the fistula is already in place. The second choice, massive hemorrhage, diverticular disease is quite likely to produce a GI hemorrhage but not so likely to produce a massive one but that's a possibility. A subphrenic abscess, well, yes, if he has

> ruptured the diverticulum, a subphrenic abscess is a possibility. Number 4, jaundice, I really don't know why they have included that choice, so I'll just ignore it. And 5, urinary obstruction, there's nothing particular here to point to this guy obstructing. So going back for the possible correct answers 1, 2 and 3. And at this point, given that he has got diverticulitis and we have evidence that he has probably dumped some nasty stuff into his peritoneal cavity, I think maybe subphrenic abscess is the choice I would go with here, so that would be choice number 3.

The first two options were eliminated using move B13 --*Eliminates an alternative in terms of likelihood* and alternative 4 was ignored (Move B6 -- *Ignores an alternative*) although an argument could be made for activity B1 -- *Reads an alternative and makes no comment.* Urinary obstruction, the fifth option, was rejected on the basis of move B15 -- *Eliminates an alternative because it does not match the data provided in the stem.* The third alternative was selected as an answer. Because the student generated answers, that is, complications, on the basis of the patient information presented in the stem and because reasons were provided for most of the discarded options, this student's overall approach to the item was also characterized by move A5 -- *Reads stem, generates answer or answer space, searches, selects, eliminates alternatives, provides rationale for most of the eliminated alternatives.*

For item 12, the most frequently used activities for disposing of the alternatives were moves B13 -- *Eliminates an alternative in terms of likelihood* (for example, "Vesico colic fistula, that's not likely.") and B16 -- *Eliminates an alternative on the basis of* *association* (for example, "Jaundice, again, jaundice would be associated with upper quadrant epigastric pain, I'd rule that out.")

In the next protocol which is for item 13, the student upon reading the stem, activated a diagnosis but offered no actions (answers) as to further investigations of the patient. Item 13 reads as follows.

> A 72-year-old man presents himself at your office complaining of weakness, easy fatigue and mild abdominal discomfort. The only finding on physical examination is pallor of skin and mucous membranes. Blood hemoglobin is 60 g/L (6.0 g/dL). A stained smear of the peripheral blood shows numerous microcytes. Rectal examination shows only an enlarged prostate.
> Your first action would be to:
> 1. Examine the bone marrow for iron stores.
> 2. Determine the serum ferritin level.
> * 3. Examine the stool for gross and occult blood.
> 4. Determine the serum iron level.
> 5. Administer blood transfusion.

The student began by reading the stem first.

> A 72-year-old man presents himself at your office complaining of weakness, easy fatigue, and mild abdominal discomfort. The only finding on physical exam is pallor of skin and mucous membranes. Blood hemoglobin is 6 g/dL. A stained smear of the peripheral blood shows numerous microcytes. Rectal exam shows an enlarged prostrate.

Having read the stem, the student restated the patient information (Move C1) by focussing on the key features and describing them in a different manner (Move C2). A diagnosis was provided for the patient's signs and symptoms (Move C4).

> Alright, he's an old man complaining of general systemic symptoms and he's definitely anemic on clinical and laboratory examination. In a guy of this age the most frequent source of blood loss is the GI tract causing an iron deficient microcytic anemia which we also have with his microcytes. His examination also shows an enlarged prostrate which in a man, 72-year-old is actually normal.

The student then turns to the alternatives.

> My first action would be ... to examine bone marrow for iron stores, I think that's a little invasive for something that is probably iron deficiency. Determine serum ferritin level, I know it will be low. Examine the stool for gross and occult blood, I think I would definitely do that. Determine the serum iron level, again, I know it would be low. Administer blood transfusion, this man doesn't need a transfusion, he's stable. What he needs is to have proof that his blood is being lost from the GI tract. I would do number 3.

Options 1, 2, and 4 were rejected using move B12 -- *Eliminates an alternative on the basis of priority* and alternative 5 was eliminated because blood transfusion is not required for a stable patient (Move B16 -- *Eliminates an alternative on the basis of association*). The keyed answer was selected because it represented the best match between the student's explanation of the patient's problem and the actions provided to resolve the patient's problem. Because the student produced a diagnosis but did not generate any answers, moves C4 -- *Activates hypothesis using the information presented in the stem* and A2 -- *Reads stem, searches for answer among alternatives, answer is triggered by alternative, eliminates remaining alternatives* typify this student's overall approach to the item.

Because item 13 requests a "first action", students used B12 -- *Eliminates an alternative on the basis of priority* to reject those options which were considered not to be appropriate or a priority.

Item 17 which appears next is the only item in this group which asks for an exception.

> The investigation of bilateral gynecomastia in a 37-year-old male should involve determining all of the following, EXCEPT:
> 1. The serum testosterone level.
> 2. A history of marijuana usage.
> 3. The serum chorionic gonadotropin level (hCG).
> 4. The patient's thyroid status.
> * 5. Ultrasound of the abdomen.

In terms of item related and successful problem solving activities, the general approach used by the students was one of reading the item and searching among the alternatives for an answer. Students did not generate answers because as was stated previously for the negatively worded items in Group One, it is more efficient for the student to search the alternatives for the exception rather than generating an exception. Students did, however, generate causes of or diagnoses which could account for the bilateral gynecomastia. In processing the alternatives, the general approach used by the students was to read each alternative and indicate whether it would or would not be one of the investigations. With respect to coding the student-option interactions for this item, options which were considered either appropriate investigations for bilateral gynecomastia or associated with a diagnosis which would have bilateral gynecomastia as one of its clinical features were labelled as activity B12 -- *Eliminates an alternative on the basis of priority* or B16 -- *Eliminates an alternative on the basis of association* respectively. Moves B12 and B16 were the most frequently used activities for rejecting options in item 17. The following think aloud protocol shows a student using some of the activities described above.

> Investigation of bilateral gynecomastia in a 37-year-old male should involve all of the following, except, so, right now I'm thinking of recurrent liver disease, could be on drugs, he could have a tumor. Those are the main ones. Let's see, all of the following, except; you would want to know: testosterone. You would want to know marijuana usage as it does contribute. hCG, well not really, number three. Number four, thyroid status, shouldn't really contribute. Ultrasound of the abdomen could be important as it could give you an idea of any intra-abdominal tumor. If he has undescended testes, but, of course, it doesn't say anything about that, so serum chorionic gonadotropin level, testicular tumors is unlikely. So testosterone level, marijuana, both are reasonable. hCG is reasonable.

> The ultrasound of the abdomen is reasonable. Thyroid status, can't say for sure but I have a feeling it's, you can have gynecomastia with thyroid disease but I'll go with number 4 because I can't think of anything better. Well, actually I'll change that, I'll go for number three as you usually don't do an hCG as an initial measure, just a follow-up. It's kind of a guess but number three.

Several diagnoses are activated by this student (Move C4). All the options were read and the student attempted to provide reasons for the inappropriate. The rationale for the investigations of serum testosterone levels, marijuana usage, and patient's thyroid status was incomplete because the student failed to indicate the information these investigations would provide. After some indecision, an answer was selected. Since the choice of answer was triggered by the alternatives, this student's overall approach to the item would be reflected in move A2 -- *Reads stem, searches for answer among alternatives, answer is triggered by alternative, eliminates remaining alternatives.*

The next question, item 18, requires students to interpret a set of blood gas values and then to select an appropriate therapy.

> A patient on a mechanical ventilator has the following blood gases:
> pO2 70 mm Hg, pCO2 40 mm Hg, [H+] 58 nmol/L (pH 7.24), oxygen saturation 91%.
> Which one of the following therapeutic maneuvers would you favor?
>
> 1. Addition of one length of dead space to ventilator circuit.
> 2. Administration of oxygen.
> 3. Administration of intravenous glucose in normal saline.
>
> * 4. Administration of intravenous sodium bicarbonate.
>
> 5. Administration of a CO2 - O2 mixture.

To answer this item, students had to determine whether the patient's blood gases were indicative of acidosis or alkalosis and whether the cause was metabolic or respiratory. In the following think aloud protocol, the student offered a correct diagnosis (Move C4) but the selection of an answer was based on a single blood gas finding.

> O.K., first of all when I look through the blood gases, I right away just make sense of the blood gases right off the bat. So, pO2 70, slightly hypoxic but nothing that's going to get you too up tight at this point. pCO2 40, he's just right. Hydrogen 58, he's acidotic and his saturation is 91% and it should be higher than that especially if he's on a ventilator. So we have some acidosis with slight hypoxia and it doesn't look like a respiratory acidosis. It's probably a slight metabolic acidosis. I'll just look at the answers. I don't even know what number one is, so I'm not even going to look at that one. Administration of oxygen, that would be possible. Administration of

> glucose, no. IV sodium bicarb, possibly depending on what is causing his metabolic acidosis but you don't want to be giving it blindly without knowing what is causing that first. So, I'm left with 2 and 4 and 5 is not correct. As always, I wish we had more information and that they would be more specific. So, well, sodium bicarb, possibly, but there is no doubt that the guy needs some oxygen, so I'll say oxygen because that is what should be taken care of before anything else.

While this student used activities such as C2 -- *Focusses on key features and defines or redescribes in a different manner*, C4 -- *Activates hypothesis using the information presented in the stem* and A2 -- *Reads stem, searches for answer among alternatives, answer is triggered by alternative, eliminates remaining alternatives*, the student's choice of answer was made using only the oxygen saturation value. Because the value of 91% is interpreted as low, the student selected an answer which would provide more oxygen to the patient. The remaining blood gas values were ignored. Option one is discarded using move B3 -- *Reads an alternative and states "Not sure" or "I don't know"*, alternative 4 was rejected on the basis of priority (Move B12. *Eliminates an alternative on the basis of priority*) while options 3 and 5 were eliminated using move B2 -- *Reads an alternative and states "No"* .

The previous think aloud protocol showed a student selecting an answer because it related to a portion of the information presented in the stem. A similar approach is used by another student responding to item 19 which reads as follows.

> A 24-year-old airline flight attendant complains of feeling tired and losing weight in spite of a good appetite. For the past year she has noticed voluminous, pale, foul-smelling stools. She recalls being told of having bowel difficulty in early childhood and of being fed a diet consisting largely of bananas.
>
> Radiological examination discloses an abnormal small bowel follow through. Biochemical analysis of the stool shows an increased amount of fat. The blood picture shows anemia.
>
> Which of the following diets would you select for this patient?
>
> * 1. Gluten free.
> 2. Lactose free.
> 3. Low fat.
> 4. Low residue.
> 5. High residue.

Upon reading the patient information in the stem, the student identified the problem as fat malabsorption (Move C4).

> 24-year-old airline flight attendant, feeling tired, losing weight in spite of a good appetite. Voluminous, foul smelling stools, makes me think of fat malabsorption. Bowel difficulty in early childhood and fed a diet consisting largely of bananas. X-ray exam abnormal bowel follow through. Biochemical analysis of the stool shows an increased amount of fat. The blood picture shows anemia. So, I'm thinking of a malabsorption syndrome but I'm not sure what kind. Which of the following diets would you select for this patient?

The diagnosis was incomplete because the student failed to integrate the additional information of "bowel difficulty in early childhood" and "a diet of bananas" into a more refined diagnosis of celiac sprue. All the alternatives were then read as a unit.

> Gluten free, lactose free, low fat, low residue, high residue. Right away, I would think number three but I'm going to go back and read through one more time, read the whole thing (Reads silently). I'm wondering what this abnormal small bowel follow through on radiological exam is. But there is increased fat in the stool so obviously, she is not absorbing fat, so I'm just going to say low fat. That is the most obvious to me.

Because the student thought the patient has a problem with absorbing fat, a diet which is low in fat is selected as an answer. Reasons were not offered as to why the remaining options were incorrect (Move B1 -- *Reads an alternative and makes no comment*).

Although the last few think aloud protocols showed students who experienced difficulty in integrating the patient clinical features presented, there were students who were able to select the keyed answer by integrating the information appropriately into a diagnosis.

The last section of Chapter IV provided several examples of how students responded to the Group Two items. The majority of these items began with a patient scenario. A review of the tabulations for the number of item related and successful problem solving activities (Appendix VIII) shows that generating answers is not a dominant activity. On the other hand, activating hypotheses during the processing of the stems is common and occurs with some consistency across students and across items.

Figure 3 shows, in the form of a flowchart, the global approaches used by the students in responding to this set of items. The double-lined arrows and the double lines indicate the pathway taken by the majority of the students. For six of the seven items, over 90% of the students began by reading the stem and activating hypotheses. In this regard, the students appear to behave like experts. They reason from data to hypothesis (forward reasoning). Move C4 (*Activates hypothesis using the information presented in the stem*) does appear to be a strong characteristic of student-item interaction. Students use other successful problem solving activities such as C1 (*Restates information presented in the item in own words*) and C2 (*Focusses on key features and defines or redescribes in a different manner*) to express the patient's clinical features as a diagnosis or a list of diagnoses (Move C4). However, the extent to which all students were able to integrate the clinical features into an accurate diagnostic category varied. This seems to suggest that these students may have some gaps in their organization of medical knowledge.

Continuing with the left hand side of the flowchart, students who activated hypotheses may also have generated answers prior to processing the alternatives. Less than 10% of the students generated answers (22 instances across the 7 items) and those who did, searched the alternatives for the generated answer or a better answer. Inappropriate options were discarded with reason. Depending upon the nature and structure of the alternatives, students used backward reasoning to discard incorrect options.

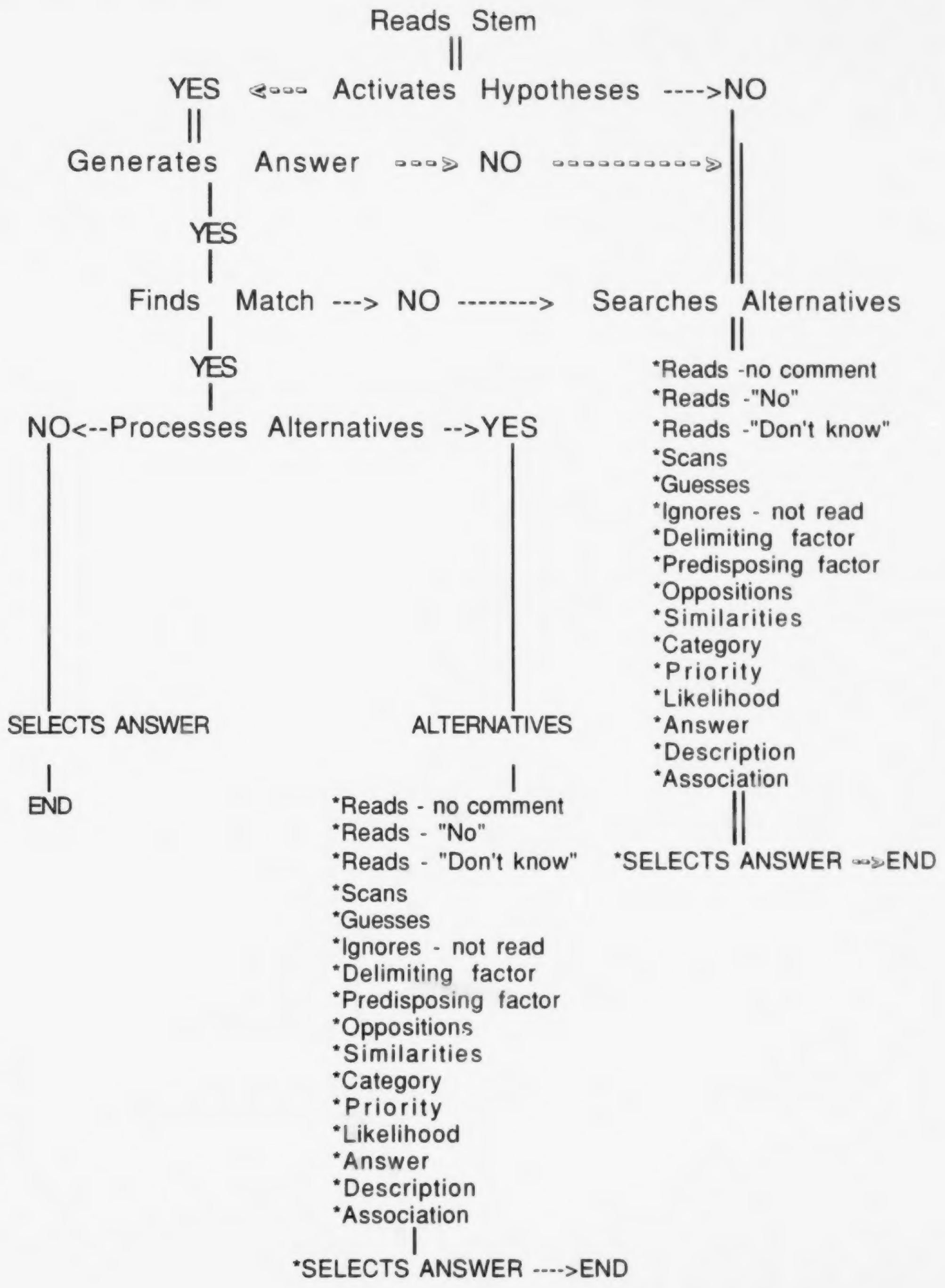
Reads Stem
YES
Activates Hypotheses
NO
Generates Answer
NO
YES
Finds Match
NO
Searches Alternatives
YES
NO
Processes Alternatives
YES
SELECTS ANSWER
END
ALTERNATIVES
*Reads - no comment
*Reads - "No"
*Reads - "Don't know"
*Scans
*Guesses
*Ignores - not read
*Delimiting factor
*Predisposing factor
*Oppositions
*Similarities
*Category
*Priority
*Likelihood
*Answer
*Description
*Association
*SELECTS ANSWER
END
*Reads -no comment
*Reads -"No"
*Reads -"Don't know"
*Scans
*Guesses
*Ignores - not read
*Delimiting factor
*Predisposing factor
*Oppositions
*Similarities
*Category
*Priority
*Likelihood
*Answer
*Description
*Association
*SELECTS ANSWER
END

FIGURE 3: Flow chart for "few or no answers generated; hypotheses activated" items

The right-hand side of the flowchart reflects the pathway taken by about 20% of the students. These students did not activate hypotheses and generate answers upon reading the stem. Typically, the entire item was read and an answer was selected by considering each of the alternatives.

Strategies for Group Two Items

From the flowchart, frequency of item-related and successful problem solving moves, and the analysis of the verbal reports, student interactions with items in this set may disclose one of three strategies.

The first strategy depicts the activities of the small number of students who neither formulated hypotheses nor generated answers (right side of Figure 3). These students read the information presented in each item and then searched the alternatives for the best answer. This is a read, search, eliminate options, and select answer type of strategy. Generally, students provided reasons for the selection of the best answer and the rejection of the incorrect options.

A second strategy characterizes the students who forwarded hypotheses and produced answers upon reading the information presented in the stem. Although the number of such students is small, their overall strategy may be described as one of reading the stem, activating hypotheses, generating answers, searching alternatives for the best answer, confirming a best answer, and discarding the incorrect options with justification.

For this set of items, the majority of the students evoked hypotheses but did not give answers. The approach of these students is captured by the third strategy. It is very similar to the second strategy and differs only in whether an answer is generated. Whereas, the second strategy shows students generating both diagnoses and answers, the protocols of students using the third strategy show that creating an internal answer space is not verbalized. The selection of an answer is determined by searching the alternatives, confirming a best answer, and providing reasons for the inappropriateness of the incorrect options.

It is for the second set of items that the majority of students show the first signs of forward reasoning. The stems for six of the seven items resemble the the kinds of cases used by Patel and Groen (1986) in their study of forward and backward reasoning. Although none of the items in this set requested a diagnosis, students upon reading the clinical features of the patients, activated hypotheses or diagnoses accounting for the patients' problems. Students reasoned in a forward fashion, going from signs and symptoms to diagnoses.

For the items in Group Two, there are 1,120 student-distractor engagements. For about 75% of these engagements, reasons reflecting knowledge and thinking were provided for rejecting incorrect alternatives. For the remaining 25%, no real reasons were given. That is, these were engagements that were coded as move B1, B2, B3, or B6.

The final section of Chapter IV deals with the third set of items.

Group Three Items

Fifteen items comprise the third group (1, 4, 5, 9 11, 14, 15, 16, 20, 21, 22, 23, 24, 27, and 28). For this item set, students activated no hypotheses but generated answers matching in form one of the item alternatives, upon reading the stem only. In terms of the successful problem solving activities identified in Chapter III move C4 (*Activates hypothesis using the information presented in the stem*) was not exercised for this item set but move C3 (*Generates answer or answer space using the information presented in the stem*) was featured for a number of items. As stated earlier, the difference between moves C3 and C4 is fuzzy and dictated by the nature of the items. To clarify this distinction between generating answers (Move C3) and activating hypotheses (Move C4), item 27 from Group Three and item 7 from the Group Two set will be used. A student's think aloud protocol for item 27 which appears below will show the generation of answers upon reading the stem only.

> A 28-year-old keypunch operator has a history of having had pneumonia four times in the past twenty years. She has had a cough "all her life" which is worse in winter. Physical examination reveals dullness, diminished breath sounds and numerous crepitations below T3 bilaterally. Her fingers are clubbed.
> She probably has:
> 1. Hypogammaglobulinemia.
> 2. Congenital heart disease.
> 3. Bronchiolitis obliterans.
> * 4. Bronchiectasis.
> 5. Cystic fibrosis.

The student's initial impression was that the patient had asthma or chronic bronchitis. Upon further interpretation of the clinical features, an additional diagnosis was activated and other diagnoses were ruled out on the basis of a single sign. This sign, clubbed fingers, was used to produce other diagnoses, two of which are identical to the answers provided in the alternatives.

> A 28-year-old keypunch operator, whatever a keypunch operator is, has a history of having had pneumonia, worse in winter. Physical examination, her fingers are clubbed. O.K., it sounds like she could have had asthma or something like that or chronic bronchitis. But when you get through the symptoms, it sounds like she has a pneumonia or something like but then, the fact that her fingers are clubbed would indicate that she doesn't have an asthma or a COPD because you don't get clubbing with either one of those. You could those with things like CF, bronchiectasis, carcinoma.

Because the student has forwarded diagnoses and since the item requests a diagnosis, the student exhibited move C3. Having developed a problem space of the patient's problems, the student reads the task set at the conclusion of the stem (what does the patient have) and then turned to the alternatives.

> They are asking what she probably has. Hypogammaglobinemia, I'm not sure. Congenital heart disease, that can give you clubbing but it sounds like she has more of a lung problem than a heart problem. Bronchiolitis obliterans is a possibility. Number 4,

> bronchiectasis, that seems to fit more or less but the fact that she has had this all her life kind of goes against that. Cystic fibrosis, that could, often presents in people that are prone to pneumonia. They have a bit of a cough all the time. The crepitations, sounds like she's got consolidation on both sides and in the bases and her fingers are clubbed. That could go along with CF or bronchiectasis, I guess, the story, but the fact that it's been so long going, I would go with CF, number 5.

The set of five options was reduced to match the generated diagnoses and the student picked CF as an answer.

To show the use of Move C4 (*Activates hypothesis using the information presented in the stem*), item 7 from the second set of items set is used. The stem in this item presents a clinical picture of a patient. However, the task for the examinee is to determine appropriate management of this patient. A diagnosis is not requested directly. Item 7 reads as follows.

> A 70-year-old man has had a smoker's cough for many years. Two weeks ago his voice became weak and hoarse, and one week ago he began to cough up 50 to 100 g of purulent sputum daily. A chest x-ray reveals a cavity containing an air-fluid level in the left upper lobe. He is afebrile.
> Which of the following would you advise?
>
> 1. Surgery with removal of the affected lobe.
> * 2. Bronchoscopy.
> 3. Sputum for acid fast bacilli culture daily for three days.
> 4. Intensive antibiotic therapy.
> 5. Postural drainage four times daily.

In the following verbal report, as the information in the stem was read, the student ruled in cancer and TB as possible diagnoses and ruled out pneumonia.

> A 70-year-old man has smoker's cough for many years. O.K., two weeks ago his voice became weak and hoarse and one week ago he began to cough up 50 - 100 grams of purulent sputum daily. The chest x-ray reveals a cavity containing an air-fluid level in the left upper lobe and he is afebrile. Alright, he is an old man who has smoked for a long time. I think of cancer right away. The hoarseness of the voice, probably due to the involvement of the recurrent laryngeal nerve secondary to the lesion that he has in his upper lobe which is hitting his left recurrent laryngeal nerve. He is afebrile, so it's probably not something like a pneumonia. Still could be something like chronic TB. Which of the following would I advise?

For this item, the task is one of selecting a course of action rather than a diagnosis. The student used the two differential diagnoses of cancer and TB to rule out less likely actions and to rule in favorable actions. The best answer was selected because the student wanted a diagnostic action.

> If I think that it's cancer, I don't think surgery will do much. Lung cancer is usually not resectable. Bronchoscopy is a good idea, it can give me a diagnosis. Sputum for AFB culture for three days. I would probably do that anyway to rule out TB because it's not unlikely. Intensive antibiotic therapy, no, because I don't think he has a pneumonia. Postural drainage four times daily, sure, I'd do that but that's not the primary problem here.

> What I would want to do is get a diagnosis here. I'd choose number 2 and do a bronchoscopy.

In terms of successful problem solving activities, this student's approach was coded as move C4 (*Activates hypothesis using the information presented in the stem*). However, had the item requested a diagnosis rather than a course of action, the student's approach would have been coded as C3 (*Generates answer or answer space using the information presented in the stem*). Because the students' reasoning in the above think aloud protocols moved from patient data to diagnosis, a claim can be made that the students were reasoning in a forward manner.

Of the fifteen items, six items presented patient scenarios and requested that students select a diagnosis. Four items required students to select a treatment or management option. In these items, the diagnosis was provided. For example, the stem for item 23 reads -- *In trigeminal neuralgia, which of the following is the treatment of choice in the early stages?*. The remaining five items dealt with general medical knowledge. For example, item 4 requested for the most frequent site of subluxation of the vertebrae in patients with rheumatoid arthritis.

In the following section, examples of students' think aloud protocols for some of the Group Three items are provided. The examples are selective because the student-item interactions and the global approaches to the items are representative of other items in the set or similar to the approaches used in responding to items in the first two groups.

The first set of four think aloud protocols are for the following item (Item 1).

> One week after an anterior myocardial infarction a 55-year-old man complains of severe pain in the left leg. The leg is cool, pale and pulseless.
> The most likely diagnosis is:
> 1. Deep venous thrombosis.
> 2. Ruptured left iliac aneurysm.
> * 3. Arterial embolism.
> 4. A-V fistula.
> 5. Arterial thrombosis.

In the first think aloud protocol, a diagnosis was generated using partial patient information. As more of the patient data was read, the initially generated answer was discarded in favor of a general answer space. Finally, the patient data were integrated to formulate another diagnosis.

> So, I'm reading, one week after an anterior myocardial infarction, a 55-year-old man complains of severe pain in the left leg. Severe pain in the left leg, right away I think D.V.T. The leg is cool, pale and pulseless. O.K., that seems different; that changes my thinking because I'm thinking instead of venous obstruction, I think cool, pale and pulseless, that sounds like an arterial obstruction and that would also fit with the severe pain in the leg. Then I'm thinking back, an anterior MI, I think if you have a severe pain in the leg in an arterial obstruction, I would think that there would be a good possibility that he could have thrown off an embolus from a mural thrombus. So, now the most likely diagnosis is: D.V.T., I wouldn't expect, I'm sort of putting that one quite down,

> scratching that one on my initial run through because of the cool, pale and pulseless. Ruptured left iliac aneurysm, that doesn't. I'm scratching that one because it makes such a nice picture of a mural thrombus and that would be coincidental to have a ruptured aneurysm. Arterial embolism, I'm going to put that one, circle that one because that's what my picture I thought of before. A-V fistula, that doesn't fit with the picture I had in my head. And arterial thrombosis, I'm going to scratch that one because arterial embolism seems a lot more likely given the anterior MI a week ago. Number 3, arterial embolism.

In the above protocol, the student exhibited some of the characteristics of successful problem solvers and experts. Upon reading part of the patient scenario, D.V.T. was generated as a possible answer. Additional patient information informed the student that the leg is cool, pale and pulseless. With this bit of information the "venous obstruction" problem space was replaced with an "arterial obstruction" problem space. D.V.T., which is a venous problem was dropped from the student's list of differential diagnoses and the patient data were integrated to form an answer of embolus. The student internalized the patient information, created a problem space, and evoked two diagnoses. One of the diagnoses was ruled out because it was not compatible with all the patient data. The early activation of diagnoses is similar to the findings reported by Elstein *et al* (1978) and Barrows *et al* (1978) in their study of problem solving. Physicians generate hypotheses early into the patient-physician interaction. Because the student was able to explain the clinical features and move from clinical features to diagnoses (forward chaining) and in the reverse direction, this student's diagnostic strategy would be categorized as diagnostic integration (Ramsden *et al*, 1989).

In terms of the examinee initiated moves identified in Chapter III, the activities of this student represent the successful problem solving moves C1 -- *Restates information presented in the stem in own words*, C2 -- *Focusses on key features and defines or redescribes in a different manner* and C3 -- *Generates answer or answer space using the information presented in the stem*. Later on in the protocol, the student provided reasons for the elimination of the wrong answers. This reflected item-related move A5 -- *Reads stem, generates answer or answer space, searches, selects, eliminates alternatives, provides rationale for most of the eliminated alternatives.*

Alternative one was discarded on the basis of association (Move B16 -- *Eliminates an alternative on the basis of association*). According to this student, if D.V.T. was to be a likely answer, then the leg would be something other than cool, pale and pulseless but because the leg is cool, pale and pulseless, the answer was not D.V.T. Options 2 and 4 were discarded using move B14 -- Eliminates *an alternative because it does not match the generated answer or the answer space* and the last option was deleted because arterial thrombosis is less likely than arterial embolism (Move B13 -- *Eliminates an alternative in terms of likelihood*).

Elimination of distractors was accomplished with backward reasoning, although for some options it was crude. The diagnosis of D.V.T. was rejected because the patient presents with signs and symptoms which are not characteristic of deep vein thrombosis. This was a direct link between the diagnosis and the data. The remaining diagnoses were rejected but not necessarily by relating diagnoses and clinical features.

The next protocol is that of a student who develops an answer of embolism (Moves A5 and C3).

> One week after an anterior MI, a 55-year-old man, severe pain in the left leg. The leg is cool, pale and pulseless. Right away from the stem, I'm thinking that we have a clinical setting in which embolic phenomena would be likely and we have the clinical presentation of a cold, pale, pulseless leg which sounds like an embolic phenomenon and I would just go immediately to the answers and look for something like an embolism.

This student integrated the patient information and formulated a diagnosis using forward reasoning. The student then went on to the alternatives and used a single move to dispose of the incorrect answers.

> And looking through the answers, 1 is not an embolism, 2 is not an embolism, 3 is, 4 and 5 neither are embolic phenomena and therefore my answer would be 3.

Move B14 -- *Eliminates an alternative because it does not match the generated answer or the answer space* was used by the student to dispense with the distractors.

The following protocol shows a student who offered a diagnosis of deep vein thrombosis upon reading all of the patient information and the diagnosis was retained as an answer after all of the alternatives have been read.

> So it says it's a 55-year-old guy who has had an anterior MI who now has severe pain in his leg and it's fairly recent, one week ago. It says the leg is cool, pale, and pulseless. So, my first thought is it's probably a D.V.T., deep vein thrombosis. So, this one is kind of easy. I'll look through it. The first one is what I thought it was. The second one, ruptured left iliac aneurysm, I guess it's possible but they say the leg is pulseless but I don't know where it's pulseless, if it's pulseless in the femorals, popliteals, if it's pulseless in the foot or not. So, arterial embolism, A-V fistula, arterial thrombosis. Now, I'm just looking down the items just to make sure, decide whether I agree with them or not, or whether I agree that I'm disagreeing with them. So arterial embolism, actually I guess that could be possible. Actually, what I was initially thinking was maybe he has thrown a clot from his heart but probably not. I think it's the D.V.T., so I think it's number 1.

In terms of item-related activities, this student's interaction with item 1 was categorized as move A4 -- *Reads stem, generates answer, searches, selects generated answer, scans or ignores remaining alternatives.* A diagnosis was activated using the patient information presented in the stem (Move C3). Because the student was moving from data to diagnosis, this would be taken as an indicator of forward reasoning. However, in this case, forward reasoning led to an incorrect diagnosis. Because the student read all of the patient information, it is impossible to determine whether the incorrect diagnosis was triggered by a specific clinical feature. For example, did "severe pain in the leg" trigger D.V.T. and this was such a powerful diagnosis that all other patient information was ignored? Or did the entire patient scenario represent for this student erroneously a wrong diagnosis because the student's knowledge base is incomplete? In terms of the diagnostic strategies proposed by Ramsden *et al* (1989), this student is using an ordering approach and the diagnosis was obtained either by exclusion or pattern matching.

Alternatives deemed inappropriate by the student were dispensed with in a superficial manner. Options 2 and 3 were considered possible but their likelihood was evaluated neither in relation to the selected answer nor the patient information presented in the stem. Options 4 and 5 were read but no comment was made about their suitability as possible answers.

In contrast to the second example, where alternatives were eliminated because they did not match the generated answer, the next protocol is that of a student who uses a pattern matching ordered approach (Ramsden *et al*, 1989) to exclude alternatives. Since an answer was not generated upon reading the patient information, the alternatives trigger an answer (Move A2 -- *Reads stem, searches for answer among alternatives, answer triggered by alternative, eliminates remaining alternatives*).

> I'm reading the question now. O.K., this is a 55-year-old man and had an MI one week ago and now has severe pain in the left leg. The leg is cool, pale and pulseless. The most likely diagnosis is and looking at the first choice

> here which says deep vein thrombosis. Deep vein thrombosis is less likely because usually you see quite a bit of swelling and it doesn't mention swelling here. So I'm going to the second one, ruptured left iliac aneurysm. If you have a ruptured aneurysm, now, I'm thinking how high the pain is and where exactly the pain is in the leg. It could be in the area of the iliac artery, I'm not sure and if you have a ruptured aneurysm, you probably don't have any pulse distal to that, so that could be. Arterial embolism, I don't think that one is that likely because it has to be a really big emboli. A-V fistula, that one somehow I don't think is likely because if you have a fistula, you have shunting of blood and it won't cause such a severe pain, I suppose, if you don't have any swelling or anything. Arterial thrombosis, again, well, I suppose it could happen. Because of the cool, pale and pulselessness of the leg, sounds like you have a decrease in arterial supply to the leg, so the first one would probably not be the answer. Second one, I would expect a lot of hemorrhage and swelling for the second one, so it's not likely. So I would choose between 3 and 5 and I think 5 is more likely. Oh!, wait a sec, this guy had MI before, so I suppose he can have emboli that originate from the heart that goes down to the leg but again, has to be a really big one. So because he's a 55-year-old too and he will probably have a lot of arthrosclerosis for developing an MI, I would probably pick number 5 as the best.

Inappropriate diagnoses were discarded by association. Usually, a single characteristic of the disease was used to eliminate an option. For example, deep vein thrombosis was considered a less likely answer because the patient does not present with swelling, although later on in the protocol, the student appeared to make the connection between cool, pale and pulselessness and an arterial supply problem.

As a result, the diagnosis of deep vein thrombosis was discarded. However, the student was unable to differentiate between an embolus and a thrombus and to integrate the patient's signs and symptoms with the potential consequences of a myocardial infarction.

Each of the alternatives served as a provisional hypothesis. For each diagnosis, the student produced an associated clinical feature and the scenario was searched for the feature. If it was absent, the diagnosis was ruled out. Backward reasoning characterizes this approach to selecting an answer.

The four examples show the different ways in which the students handled the same item. Upon reading the patient information presented in the stem, diagnoses were generated by the first three students. The first student altered the problem space from venous to arterial as more of the patient information was read. Clinical features of the patient were integrated and the student was able to move from a diagnosis of deep vein thrombosis to a correct diagnosis of arterial embolism. The second student integrated the patient data into a single problem space (embolic phenomenon) and then generated the correct diagnosis. Forward reasoning was also a characteristic of the third student's protocol, however, a wrong diagnosis was generated after reading the patient information and was retained after all of the alternatives were read. The last student did not generate an answer upon reading the patient information. Upon reading the alternatives, the student was able to associate the patient's clinical features with an arterial supply problem. This action reduced the choice of answers to options 3 and

5 and although the student had a plausible explanation for option 3 being correct, the student chose instead to go with option 5, an incorrect choice.

The selection of incorrect answers by two of the four students raises the question of what misconceptions students had about the knowledge being tested by the item. It is likewise a question of how knowledge was organized and what errors in reasoning led to a wrong diagnosis or an incorrect answer. Although such questions were not the focus of this study, it is evident from the protocols of the two students who answered item 1 incorrectly that misconceptions appear at different levels. In one case, the student read the patient information and generated a wrong answer (forward reasoning) but was not able to differentiate between a venous obstruction and an arterial obstruction problem. The second student did not generate any answers but after reading the entire item was able to identify the problem as one of decrease in arterial flow. The student was then able to discard diagnoses involving the venous system, however the student was not able to differentiate between an embolus and a thrombus, both of which are arterial problems.

For the other five items which requested a diagnosis, students approached these items in ways similar to the example think aloud protocols presented above for item 1. At least half of the students generated diagnoses (Moves A3, A4, or A5 and C3) on the basis of the information presented in the stem only. These students were able to develop for each item a cognitive representation (problem or answer space) of the patient's presenting problems. However, developing a problem space or generating an answer did not necessarily lead to the selection of the correct diagnosis as is shown in the following protocol for item 5.

> A 22-year-old male has had intermittent low back pain for four years. It does not seem to be related to anything in particular, usually lasts 2-3 days. He is now concerned because his right eye is red and painful and the vision is decreasing. So, to me, I start thinking of the spondyloarthropathies, something like Reiter's syndrome or ankylosing spondylitis or something like that. So then, I go down and the choices are rheumatoid arthritis, ankylosing spondylitis, TB or sarcoidosis. I don't consider those but my two choices are between AS and Reiter's syndrome and I'm just trying to remember. I think that it's Reiter's syndrome that's mostly associated with eye findings, so I think I would pick number 5.

Although two diagnoses were generated, one of which is correct, the student made an incorrect choice on what appears to be an incomplete interpretation of the patient data. The student used a single clinical feature (eye findings) to select a diagnosis. In terms of the diagnostic strategies presented by Ramsden *et al* (1989), this student would be using an ordered pattern matching approach. A short associative link between eye findings and the diagnosis was made. While it appears that the student integrated the patient's clinical features to develop a problem space, the student was unable to tease out the salient difference in the clinical features of a patient with ankylosing spondylitis and one with Reiter's syndrome. Both patients would present with some form of arthritis and eye problems, however, a patient with Reiter's syndrome would also have urethritis. Reiter's syndrome consists of the triad of arthritis, conjunctivitis and urethritis whereas urethritis is not associated with ankylosing spondylitis.

In addition to failing to develop an adequate cognitive representation of the patient's problem and to trigger the correct answer, students sometimes adhered to a diagnosis, usually evoked on the basis of one or two clinical findings. Such diagnoses were activated early in the reading of stem. Although there was additional evidence to suggest another diagnosis, the initial diagnosis was retained as the final answer. Kassirer and Kopelman (1989) refer to this adherence to an inappropriate diagnosis in spite of sufficient information for another diagnosis as premature closure. An example of a student using premature closure is presented for item 14 which reads as follows.

> A 56-year-old man presents to his doctor with a month history of intermittent right facial pain. On examination he is found to have a diminished corneal reflex and slight hearing defect on the right.
> The diagnosis is:
> 1. Right cerebral tumor.
> 2. Trigeminal neuralgia.
> 3. Otitis media.
> * 4. Acoustic neuroma.
> 5. Multiple sclerosis.

The student generated a diagnosis (Move C3 -- *Generates answer or answer space using the information presented in the stem*) using the clinical feature of intermittent right facial pain. Upon reading the remainder of the stem, the student stated that the additional patient information was not consistent with the generated diagnosis. No attempt was made to refine the problem space and the student then proceeded to the alternatives.

> Fifty-six-year-old man, intermittent right facial pain. That's trigeminal neuralgia that comes to mind right there. Diminished corneal reflex and slight hearing defect on the right. Those aren't completely, those aren't entirely consistent with trigeminal neuralgia, so I'll take a look at the answers. Right cerebral tumor, cerebral tumor, no, that doesn't make sense. It would have to be a brain stem lesion. Trigeminal neuralgia, it's possible. Otitis media doesn't make any sense really at all. Slight hearing defect, diminished corneal reflex. Why is his corneal reflex diminished if he has otitis media. Acoustic neuroma, that's a possibility but I wouldn't go with that. It would affect his hearing more than anything else. Multiple sclerosis, that is a possibility but it seems to fit in more likely with trigeminal neuralgia -- 2, because it came to mind first, I guess I would say that.

Alternatives 1, 3 and 4 were discarded using move B16 -- *Eliminates an alternative on the basis of association* while move B13 -- *Eliminates an alternative in terms of likelihood* was used to eliminate the fifth option. In terms of item-related activities, this student's overall approach to item 14 was characterized by move A5 -- *Reads stem, generates answer or answer space, searches, selects, eliminates alternatives, provides rationale for most of the eliminated alternatives.*

After processing the alternatives, the student retained the diagnosis generated initially (trigeminal neuralgia) and which failed to explain all of the clinical features. Options appeared to be eliminated using a single clinical feature (Ordering approach using exclusion in terms of Ramsden *et al* (1989)) and although the clinical finding may be a characteristic of the diagnosis, it may not necessarily be the differentiating characteristic. In the case of the think aloud protocol above, the student stated that if acoustic neuroma was to be considered as a diagnosis then the patient's hearing would be affected to a greater degree. The student was neither able to explain all the clinical features to defend the choice of diagnosis nor able to rule out some of the competing diagnoses by comparing the clinical features presented in the stem with the features of the diagnoses presented in the alternatives. Premature closure appeared because the student seemed unable to interpret the clinical features appropriately.

Nineteen students produced an answer upon reading the information presented in the stem. Of these, 15 students gave a diagnosis of trigeminal neuralgia after reading the first sentence and of the fifteen students, eight retained the diagnosis after reading the additional patient information and the alternatives (premature closure). While these students were able to display forward reasoning for part of the patient information, these students were not able to integrate all of the patient findings. These students were not behaving like true experts. Their overall diagnostic strategies would be labelled as ordered (Ramsden *et al*, 1989) and the selection of a diagnosis was either by exclusion or pattern matching. The remaining seven students who also gave a diagnosis of trigeminal neuralgia after reading the first sentence, selected an answer other than trigeminal neuralgia after reading the entire item. These students were able to correct their initial impression of the patient's problem. Of the nineteen students, three produced a diagnosis of trigeminal neuralgia after reading the entire stem. Two of these three students retained the initially generated diagnosis of trigeminal neuralgia after processing the alternatives.

In contrast to the above student's interaction with item 14, the following think aloud protocol shows a student who does not close the problem prematurely. Instead, after the information presented in the stem was read, the patient's clinical features were explained in terms of nerve involvement. Having identified the nerves, the student generated an answer of CPA tumor.

> A 56-year-old man presents to his doctor with a one-month history of intermittent right facial pain. On examination he is found to have a diminished corneal reflex and slight hearing defect on the right. O.K., before I look at the answers. A right facial pain, because it's limited to the right side suggests involvement of a nerve as opposed to muscle or trauma or anything like that. There is nerve involvement here and that's documented again by diminished corneal reflex which suggests that either his fifth nerve, the sensory component of the fifth nerve, or the seventh nerve is involved. The slight hearing defect on the right suggests involvement of the eighth nerve and since the eighth nerve runs along with the seventh nerve, I'm thinking of something that would involve both the seventh and eighth cranial nerves and the thing that comes to mind is a CPA tumor. And if I go

> to the answers, I see CPA tumor under number 4, acoustic neuroma, and just from the fact that this is what is normally presented, I would go with acoustic neuroma.

In interacting with item 14, this student uses the successful problem solving moves C1 -- *Restates information presented in the stem in own words*, C2 -- *Focusses on key features and defines or redescribes in a different manner* and C3 -- *Generates answer or answer space using the information presented in the stem.* Because no overt evidence was provided in processing the alternatives (Move B6 -- *Ignores an alternative*), the student's interaction with the item was also coded as move A4 -- *Reads stem, generates answer, searches, selects generated answer, scans or ignores remaining alternatives.* This student behaved like an expert. The approach was structured and the diagnosis was reached through diagnostic integration (Ramsden et al, 1989). All the clinical features were integrated and explained without introducing new data. Because the pattern of reasoning is from data to diagnosis, this student used forward chaining.

For some of the items, students offered either a very brief or no explanation of the information contained in the stem. No overt evidence was provided for the use of moves C1 and C2, however, answers were generated (Move C3) and once an answer was selected, the remaining alternatives were either ignored (Move B6) or confirmed as inappropriate. Think-aloud protocols showing some of these features are presented for items 4, 11, 20 and 23. Item 4 reads as follows.

> The most frequent site of subluxation of the vertebrae in patients with rheumatoid arthritis is:
>
> * 1. C1-2
> 2. C5-6
> 3. L4-5
> 4. L5-S1
> 5. L5-S2.

In the first example, the student forwarded an answer (C1 C2) and qualified this answer by stating that the site is in the neck.

> The most frequent site of subluxation of the vertebrae in patients with rheumatoid arthritis is, right away, for some reason, I know it's C1 C2, before I even look at the answers. I recently read that it's neck subluxation, so looking at those answers, I'll rule out answers 3, 4 and 5 right away and between 1 and 2 and I just have a feeling it's 1. The answer is number 1. Actually, I know the answer is number 1, I say with confidence, ready to get the question wrong when it comes back to me.

Because alternatives 3, 4 and 5 are lumbar rather than the cervical (neck) vertebrae, they were eliminated. No comment was made about option 2 (Move B6 -- *Ignores an alternative*) and options 3, 4 and 5 were eliminated using move B11 -- *Eliminates an alternative using category or class.* These two moves were used frequently to discard the options in item 4. Upon reading the stem only or the entire item, the majority of the students generated either a specific answer (for example, C5-6) or a problem space (cervical or lumbar). Options which were not related to the specific answer or did not belong to the problem space were not considered. In the following example, the student stated that the subluxation occurs at the cervical level. Only the first two options were considered as possible answers.

> The most frequent site of subluxation of the vertebrae, subluxation in vertebrae in patients with rheumatoid arthritis is at the cervical level, C1-2, C5-6. I think C1 C2, yes, C1 C2.

After the first option was selected as an answer, the student was probed with the question, "Did you do anything with the others?" and the reply was "No, I know from reading it's cervical, so I just looked at 1 and 2." The student appears to be confident with the choice.

Similar student-item interactions were noted for item 11 which reads as follows.

> At present, the agent of choice for MRSA (Methicillin resistant Staphylococcus aureus) is:
> 1. Cloxacillin.
> * 2. Vancomycin.
> 3. Ceftazidime.
> 4. Erythromycin.
> 5. Cephalothin.

In the think aloud protocols that follow, students produced the correct answer without referring to the alternatives (Successful problem solving move C3 -- *Generates answer or answer space using the information presented in the stem*). In the first three examples, distractors were ignored (Item-related move A4 --*Reads stem, generates answer, searches, selects generated answer, scans or* *ignores remaining alternatives* and disposition of alternatives move B6 -- *Ignores an alternative*).

> O.K., so the agent of choice for Methicillin resistant staph. Without looking at the answers, I think it's Vancomycin. So, I would go with number 2 and basically ignore the rest.
>
> Agent of choice for Methicillin resistant Staph aureus is Vancomycin. I have to look for Vancomycin. I see it, number 2. That's my answer. I won't even look at the other ones.
>
> Agent of choice for, is Vancomycin, just from memory.

Some students generated similar protocols for item 20 which requested the most definitive diagnostic test for pulmonary embolism with or without infarction and for item 23 which dealt with the treatment of choice for trigeminal neuralgia. An example for each item is presented below.

> The most definitive diagnostic test for pulmonary embolism with or without infarction is, O.K., I know that you do a ventilation-perfusion scan but it can sometimes not give you a definitive diagnosis, so, even though it's the first thing you do, you'd probably do an angiography too. So, I'm going to look in the answers to see if there is angiography. Pulmonary angiogram, number 5, so I'll choose that.
>
> In trigeminal neuralgia, which of the following is the treatment of choice? Since I have done some neurology, Tegretol, number 1, don't even have to think about it.

In these examples, students were able to produce answers on the basis of the information presented in the stem. In this study, generating an answer was taken as evidence for forward reasoning. However, the students' responses in the examples above appear to be automatic and based on the recall of memorized facts. According to Smith (1991), a problem is a task that requires analysis and reasoning toward a solution. To resolve a problem, more than recognition or recall is required. Using the descriptions provided by Smith, the verbal reports presented above would not be considered as exemplars of forward reasoning. But the degree of analysis and reasoning required in reaching a solution would be influenced by the cognitive demands made by the item on the student's declarative and procedural knowledge. For some students these items made little demand and could be answered on the basis of declarative knowledge. These items may be viewed as similar to the demands imposed by asking an adult "How much is 2 plus 2?"

Other students produced different verbal reports for items 4, 11, 20 and 23. The next example shows a student who generated the correct answer for item 11 and in spite of stating that there is no need to check the remaining alternatives, the student read the other options and verified that option 2 is indeed the best answer.

> I'll read it out loud. O.K., MRSA, I know for a ... I don't even have to look at the answers. I think about it, first choice is Vancomycin, if it's on there, yep, number 2. I'm not even going to look at the other ones but I will any ways because I'm paranoid and we're taught to be paranoid. So,

> I'll look at them. Clox, no. Ceftazidime, no. Erythromycin, no. Nope, yep, it's Vanocmycin.

In the next two verbal reports for items 11 and 20 respectively, students do not generate answers. Instead the entire item is read and the alternatives serve as a trigger to an answer (Move A2 -- *Reads stem, searches for answer among alternatives, answer triggered by alternative, eliminates remaining alternatives*).

> At present, the agent for choice for MRSA, Methicillin resistant Staph aureus, is; O.K., now look through these and I'm not sure. Clox, I know you use in Staph. Vancomycin, I know you use in Staph. Ceftazidime, it's primary use is in pseudomonas, that's its big claim to fame. Erythromycin, I don't think of that one as being associated with, that's not a first line for Staph. Cephalothin, I don't know the primary of what Cephalothin is used for. Ceftazidime, I'm scratching because it's primary association is pseudomonas. Erythromycin, I don't think of it as being associated with Staph particularly. Cephalothin, I don't know and because I have two that I am comfortable with Staph, I'm going to scratch that one as being likely. Cloxacillin, when I've used it, I think of it as being a first line drug for Staph and so I'm thinking if it's Methicillin resistant, there is a good chance that the Cloxacillin and the Methicillin can be thought of as having the same sort of resistance. So, I'm thinking of Vancomycin, so, I'm going to go with Vancomycin, number 2.

A variety of reasons were provided for the incorrect choices. Option 1 was discarded using move B10 -- *Eliminates an alternative on* *basis of similarities*; move B16 -- *Eliminates an alternative on the basis of association* was used to eliminate option 3 and the last two options were dropped using move B3 -- *Reads an alternative and states "Not sure" or "I don't know"*. In the next think aloud protocol, the distractors are likewise discarded with reason.

> The most definitive, without infarction is. So, they're asking what test you would use to tell that there is a pulmonary embolus. Perfusion lung scan, that's a possibility because that would show areas of decreased perfusion but decreased perfusion could be due to dead space as well as pulmonary embolus so that's not that likely. Number 2, VQ ventilation perfusion lung scan, I think is probably the right answer. Number 3, a decreased arterial pCO2, could be caused by so many things; that's not definitive at all. An increased AA D O2, that's not definitive for pulmonary embolism. Number 5, pulmonary angiogram, I suppose that could show, that's a possibility. So my choices now would be between number 2 and 5. Number 2, ventilation perfusion scan, I'm trying to remember; they do the ventilation part first and if there is abnormalities in the ventilation, they don't bother with the perfusion because the perfusion will be altered to the areas without ventilation in them, the dead spaces. I think, it's probably a VQ lung scan, a VQ mismatch is indicative of pulmonary embolism I think, or gives a high probability of pulmonary embolism. I'm not sure it's definitive. Pulmonary angiogram would probably definitely give you evidence of one. I'm not sure how often they would use it as a test for one but I guess if they want something totally definitive, that will tell you what it is, number 5.

The student used move B13 -- *Eliminates an alternative in terms of likelihood* to eliminate option 1 and the other three distractors that were considered as not being definitive tests were discarded using move B9 -- *Eliminates an alternative using oppositions.*

In general, the verbal reports for Group Three items, in comparison to the two other groups of items, were characterized by the frequent use of answer production upon reading the stem only (Appendix VIII). This is reflected in the frequency with which students use item-related activities A3 -- *Reads stem, generates answer, searches for answer, alternative triggers different answer, discards initial answer and eliminates other alternatives,* A4 -- *Reads stem, generates answer, searches, selects generated answer, scans or ignores remaining alternatives,* and A5 -- *Reads stem, generates answer or answer space, searches, selects, eliminates alternatives, provides rationale for most of the eliminated alternatives.* The total use of these moves is reflected by the successful problem solving move C3 -- *Generates answer or answer space using the information presented in the stem.* The number of times move C3 was used ranges from 9 times for each of items 9, 23, and 28 to 32 times for item 1. In addition to using move C3, students made use of moves C1 -- *Restates information presented in the stem in own words* and C2 -- *Focusses on key features and defines or redescribes in a different manner.*

Those students who did not generate answers on the basis of the information presented in the stem, used item-related activities A1 -- *Reads stem and all alternatives* and A2 -- *Reads stem, searches for answer among alternatives, answer triggered by* *alternative, eliminates remaining alternatives* and the successful problem solving moves C1 and C2 to answer items. For these students, the alternatives triggered the selection of an answer.

Global approaches to answering the Group Three items are shown in Figure 4.

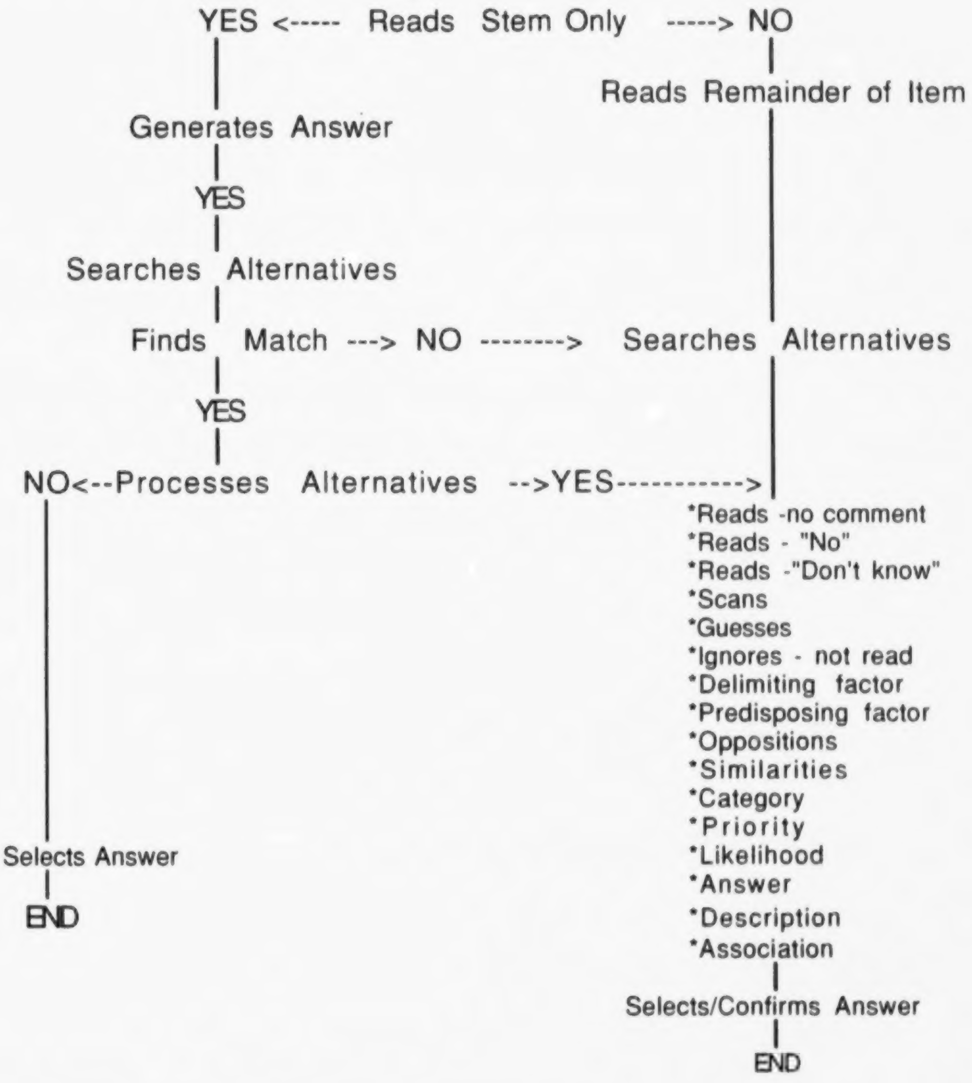
YES <----- Reads Stem Only -----> NO
Reads Remainder of Item
Generates Answer
YES
Searches Alternatives
Finds Match ---> NO --------> Searches Alternatives
YES
NO<--Processes Alternatives -->YES----------->
*Reads -no comment
*Reads - "No"
*Reads -"Don't know"
*Scans
*Guesses
*Ignores - not read
*Delimiting factor
*Predisposing factor
*Oppositions
*Similarities
*Category
*Priority
*Likelihood
*Answer
*Description
*Association
Selects Answer
END
Selects/Confirms Answer
END

FIGURE 4: Flow chart for " answers generated; few or no hypotheses activated" items

Two major pathways constitute the flowchart. The first decision point is the stem and whether an answer matching in form one of the alternatives is produced. Given that for this set of items, there is a total of 600 answers (15 items x 40 students), 296 answers were produced upon reading of the stem only. These students follow the left hand side of the flowchart. On generating an answer, the student then engages in a search of the alternatives for a match. The alternatives constitute a second point of decision If a suitable match is found the student may bypass the processing of the options entirely. Of the 296 instances where answers were produced, there were 70 recordings of move A4 -- *Reads stem, generates answer, searches, selects generated answer, scans or ignores remaining alternatives* where no reasons were given for the the elimination of incorrect answers. On the other hand, a student may, for a variety of reasons, process the alternatives to verify that the generated answer is indeed the best answer. There were 137 recordings of this activity (Move A5). Alternatives are also considered if the generated answer is outside the answer space encompassed by the alternatives. There were 89 instances in which students generated an answer but upon reading the alternatives, replaced the generated answer with one of the alternatives. The pathway concludes with the selection of an answer.

The right hand side of the flowchart represents the approach followed by those students who do not advance an answer upon reading the stem only. Instead these students read the stem and then proceed to the alternatives. Options may be read as a unit and then processed individually or the student may choose to deal with the options, usually, in a sequential manner. As each option is read, a reason is provided for its potential as an appropriate answer. The pathway concludes with the selection of an answer.

Strategies for Group Three Items

The analysis of the Group Three items reinforces the findings reported for the other two item sets. Students use a variety of approaches in their encounters with multiple choice items. For this set of items, the examinee-initiated activities can be combined to produce one of three general strategies. The first strategy, which was also used by students for the other item sets, applies to situations where answers are not advanced after reading the stem. In this strategy students read the stem and then proceed to one of two approaches in dealing with the alternatives. This strategy is represented by the right hand side of Figure 4. The alternatives may be read as a unit after which a more deliberate processing of each alternative is undertaken or the alternatives are processed in a sequential fashion. Normally, the first alternative is read and a rationale is presented by the student for its suitability as an answer. This process continues for the remaining alternatives and concludes with the student selecting an answer. Reasons are provided for most or all of the alternatives discarded as incorrect answers.

The second and third strategies, which include the generation of answers upon reading the stem only, are represented by the left hand side of Figure 4. In the second strategy, students read the stem, generate an answer, and then go on to the alternatives in search of the generated answer or a better answer. Incorrect alternatives are rejected with reason. The third strategy, an offshoot of the second strategy, consists of students reading the stem, generating an answer, searching the alternatives for a match, finding the match and then ignoring the distractors. That is, once the generated answer is found in the alternatives, the student does not provide any overt evidence for discarding the wrong answers. This is a read stem, generate answer, find answer and move on to the next item type of strategy.

Coupled with each of these strategies is a component made up of successful problem solving moves. Upon reading the stem or the entire item students may restate the item in their own words, provide a more in depth description by referring to etiology and pathophysiology or developing categories of diseases. If the students produce these descriptions by building on the information presented in the stem or entire item, an argument can be made that forward reasoning is being used. However, this is only a partial representation of how the items are answered. As stated earlier, incorrect options are usually rejected with reason. Each alternative serves as a provisional hypothesis and its sufficiency as an answer may be determined by referring either to the information presented in the stem or to the student's own internalization of the problem. In such instances, the discarding of alternatives is done through backward reasoning.

Summary

The verbal reports of forty third- and fourth- year medical students on thirty multiple choice items were grouped according to the item-related moves A2 (*Reads stem, searches for answer among alternatives, answer triggered by alternative, eliminates remaining alternatives*), A3 (*Reads stem, generates answer, searches for answer, alternative triggers different answer, discards initial answer and eliminates other alternatives*), A4 (*Reads stem, generates answer, searches, selects generated answer, scans or ignores remaining alternatives*), and A5 (*Reads stem, generates answer or answer space, searches, selects, eliminates alternatives, provides rationale for most of the eliminated alternatives*) and the successful problem solving moves C3 (*Generates answer or answer space using the information presented in the stem*) and C4 (*Activates hypothesis using the information presented in the stem*). Moves A3, A4, and A5 are reflected in move C3. Three groupings of items emerged. The first group consisted of those items for which students, upon reading the stem, generated few or no answers and evoked no hypotheses. Activating hypotheses was a strong characteristic of the verbal reports produced for the second group of items. In this set of items, students generated few or no answers upon reading the stem. The third group consisted of those items for which students did not summon hypotheses but did extract answers upon reading the stem.

In Chapter V, these three groupings will be examined from the broad perspective of clinical problem solving and some suggestions for item development and future research will be made.

V. DISCUSSION, IMPLICATIONS AND REFLECTIONS OF MULTIPLE CHOICE ITEMS AND CLINICAL PROBLEM SOLVING

At the outset it is important to note that the most dominant impression that is gained from examining the data is the enormous variation that arises from different individual encounters with different items. Each examinee brings to the exercise his or her own organization of knowledge, medical experience, confidence in the knowledge of the facts, approach to reading, and response to the "think aloud" task. When this is coupled with the semantic, syntactic, and propositional characteristics of the items, it is very difficult to predict in advance the sequence of the thought process that will occur.

Nevertheless, it is helpful to consider the broad classes of examinee-item engagements that did occur. It appeared most reasonable to cluster these engagements according to items, rather than examinees. Several reasons account for this. First, certain semantic features of the items seemed to dominate or constrain the strategies. For example, items presenting patient clinical features usually elicited diagnoses from the students, even though diagnoses were not the main point of the items. On the other hand, students did not generate answer spaces for items in which the stems were stated negatively or were unfocussed. Second, the examinees were not experts, at least in all domains, and so they were forced to reason from specific features of each item.

It seemed that the item clusters were relatively unrelated to statistical item difficulty, the associative strengths of the distractors, or to the loosely defined levels of Bloom's Taxonomy. If expertise in this context is indicated by the amount or pervasiveness of forward reasoning shown in the responses to the items, then roughly speaking, the clusters fell along a continuum of expertise. For most of the items in the first cluster, students offered few answers or hypotheses upon reading the stems. Forward reasoning was seldom used in the student-item interactions. Instead, students read the entire item, processed the information and then, either confirmed or rejected each alternative using backward reasoning.

For the second group of items, only a small number of students generated answer spaces. However, the majority activated hypotheses. The difference between generating answers and activating hypotheses is not always clear. Generating answers means that the elements of the generated answer or problem space approximately match the alternatives in content and form. Activating hypotheses could be thought of as being prior to generating answers. Students formulated diagnostic descriptions of the patient's clinical features, even though none of the items in the second cluster requested a diagnosis. Instead some other aspect, such as management and treatment, was required. What this shows is that prior to the task posed by the item, students compile and integrate the patient's clinical features into diagnostic labels. Because the reasoning proceeds from clinical features (signs, symptoms, laboratory results) to diagnosis, reasoning appears to be predominantly forward. Generating answers and activating hypotheses occur upon reading the information presented in the stem and prior to the processing of the alternatives. Most of the items in the second set were based on clinical scenarios and, in a limited way, bear some resemblance to the case descriptions used by researchers such as Patel and Groen (1986), Lemieux and Bordage (1986), Ramsden *et al* (1989), and Schmidt *et al* (1990).

In the third cluster of items, students generated answers but activated few hypotheses. That is, the elements of the answer space resembled, in content and form, the set of alternatives. Some of the items in this set were based on clinical scenarios and requested a diagnosis while other items required students to select an investigative procedure for a specific condition.

For all item sets, backward reasoning was evidenced in the elimination of alternatives. That is, the suitability of each alternative as a plausible answer is matched against the information presented in the stem. Not only do the format features of the multiple choice item make it prudent for students to use backward reasoning to be sure that other alternatives are not as good as the selected answer but the very nature and practice of medicine requires that, at least, all the common possibilities be considered.

If forward reasoning is used as the sole criterion for expertise, then for some items, students are behaving like experts. That is, for many student-item interactions, the multiple choice item does elicit very complex reasoning that is clinically relevant. Many of the thirty multiple choice items used in this study require more than just reading the stem and recognizing the correct answer from the list of alternatives. Hence, criticisms of the format are not entirely valid. According to Snow (1993), viewing multiple choice as "multiple guess" items which require nothing more than rote response regurgitation and the selection of the correct answer may be so only in the eye of the critic; the test-taker may view the item very differently.

In addition to exhibiting instances of forward and backward reasoning, the students displayed several other characteristics of expertise and successful problem solving. Students applied their knowledge, internalized the information presented in the items, created problem spaces, and checked the plausibility of the selected answer. On the other hand, experts command a large amount of information and concepts, organize their knowledge into richly connected schemas and illness scripts, and perform tasks rapidly and efficiently. These aspects were less obvious in the reasoning students displayed in their interactions with the items. The focus of the present study was on the identification of strategies that students actually used in responding to multiple choice items. Whether the student-item interactions were manifestations of the deeper qualities of clinical reasoning characteristic of expertise such as richness of scripts, chunking, and automaticity was not investigated. The evidence was not available in the think aloud protocols.

Because the purpose of this study was to present a preliminary and holistic description of what transpires when students respond to multiple choice items, the focus was on mapping the field. The results direct us to select items carefully for future investigations. Items with unfocussed and negatively worded stems do not appear to elicit successful problem solving skills such as generation of problem and answer spaces and forward reasoning. Other features of items such as stems that seek diagnoses, given certain patient characteristics encourage forward reasoning. To understand how items elicit important aspects of critical thought requires the creation of items with appropriate syntactic and semantic forms. Perhaps what is required for the next research in this area, is to have item writers produce reasoning scripts for each newly generated item. Such a request could accomplish at least three things. First, the new items along with the accompanying scripts matched to the appropriate level of training might be more reflective of the declarative and procedural knowledge that students should possess. Second, the reasoning scripts would act as a check for determining whether the stem presents all the necessary information for the item to be answered correctly. Finally, the scripts could force item developers to develop more meaningful distractors reflecting the misconceptions about the particular content. In other words, if the responses to test items are to be taken as a signs and samples of behavior, then the item must have substantive psychological meaning. Having item writers produce reasoning scripts for items could enhance the psychological meaningfulness of items. Items developed in the way suggested may be more appropriate for study than some set of arbitrarily selected items from an existing pool.

Methods such as propositional analysis (Patel & Groen, 1986) and structural semantics (Lemieux & Bordage, 1986) could be used to analyze the features of the verbal protocols. These methods can most profitably be used after the demand characteristics of the items are better understood. The present research indicates that differences in levels of expertise interact with the nature of the task, the subtleties that distinguish the alternatives, and the format to influence the examinees' reasoning. The next stage of the research is to focus on a relatively uniform set of items and carry out a deeper linguistic analysis.

Although researchers investigating the structure of clinical reasoning tend to have a preference for either propositional analysis (Patel & Groen, 1993) or for structural semantics (Lemieux & Bordage, 1993), future research should consider integrating the two methods. In reviewing the think aloud protocols, there were instances where students used a mixture of elements from propositional analysis (if-then causal rules) and from structural semantic analysis (binary oppositions) to internalize the meaning of information presented in the item. For some items, selection of the correct answer and the elimination of distractors were dealt with primarily from a structural semantic perspective. Hence, using a single method would not likely capture all that transpires in these student-item engagements.

Several implications emerge for testing. Cognitive psychology views students as active information processors and constructors of knowledge who accumulate declarative knowledge and integrate the newly acquired facts and concepts into existing knowledge structures. By automating procedures and chunking information, existing schema and scripts are reconfigured (Snow & Lohman, 1989;

Tittle, Hecht, & Moore, 1993; Mislevy, 1993) and expertise is said to be acquired.

Considering the advances in cognitive psychology, it seems reasonable to adopt the theory that cognitive skill acquisition consists of declarative, knowledge compilation, and procedural stages (Royer, Cisero, & Carlo, 1993). Coupling the cognitive psychology perspective with the variability witnessed in the student-item interactions would suggest that designing items according to the various levels of Bloom's Taxonomy is not useful. At best the Taxonomy should be viewed as a description of the intentions of the item writer. It appears to have little empirical relationship to the thought processes of examinees. Analysis of the think aloud protocols revealed that for a single item, students' thought processes crossed several or all of the major levels of Bloom's Taxonomy. At the same time, there were items which were answered quickly and without much overt evidence for reasoning.

The problem from the test developer's point of view is that it is unknown in advance what "illness scripts" will be elicited by the students because their level of expertise is not necessarily high. Indeed, it could be argued that items should be aimed at some hypothetical level of expertise that challenges the examinee. As noted earlier, part of this problem may be resolved by requesting item writers to think in terms of scripts, forward chains, and appropriate hypothetico-deductive sequences. Each distractor, representing known misconceptions, could be linked by a hypothetico-deductive reason and the keyed answer by an appropriate forward chain. To aid with the development of distractors, item writers should think in terms of cognitive errors that students are likely to commit. This suggestion stems from the work of Kassirer and Kopelman (1989). In reviewing the think aloud protocols of physicians, the authors identified four classes of errors related to faults in triggering hypotheses, context formulation, data gathering, and verification. Some of these errors, such as faulty triggering of hypotheses (diagnoses) and faulty verification as a result of premature closure, could be identified at least in some form in the students' interaction with the multiple choice items used in this thesis.

In addition to asking item writers to provide reasoning scripts for items, items should be designed so as to be free of technical flaws. For example, stems which consist of single words or simple phrases or stems which are stated negatively do not appear to facilitate the understanding of problem solving.

At the very least, for high stakes examinations, part of the try-out process could collect some limited think aloud protocols from examinees at the appropriate level of expertise. An efficient structural and propositional analysis could then be used to feed information back to the item writers to guide their revisions. By coupling the feedback with the suggestion to have item writers think in terms of illness scripts and forward and backward chaining, an argument can be made for the validity of the test item as a stimulus feature based on the cognitive processes appropriate for reaching an answer. Such a process would also have an immediate implication for standard setting on high stakes exams. Using this framework, setting a passing score would be viewed as an extension of validity and hence a cognitive, rather than a demographic, concept (Maguire, Skakun, & Harley, 1992; Maguire, Hattie & Haig,1993).

Finally, the think aloud procedure could be used to investigate differences between item formats (Traub, 1993). For example, multiple choice items designed according to the suggestions presented earlier could be converted to constructed-response items. Applying structural and propositional analyses to the think aloud protocols of the two item sets would help determine whether examinees use different cognitive processes when responding to multiple choice and constructed-response items.

References

Afflerbach, P., & Johnston, P. (1984). On the use of verbal reports in reading research. *Journal of Reading Behaviors*, *16,* 307-322.

Anzai, Y. (1991). Learning and use of representations for physics expertise. In K.A. Ericsson and J. Smith (Eds.), *Toward a general theory of expertise* (pp. 64-92). New York. NY.: Cambridge University Press.

Barrows, H. S. (1986). The scope of clinical education. *Journal of Medical Education, 61* (9), 23-33.

Barrows, H. S., Feightner, J. W., Neufeld, V. R., & Norman, G. R. (1978). Analysis of the clinical methods of medical students and physicians. Report submitted to the Province of Ontario Department of Health and Physician's Services Inc. Foundation.

Barrows, H. S., & Feltovich, P. J. (1987). The clinical reasoning process. *Medical Education, 21,* 86-91.

Barrows, H. S., & Tamblyn, R. M. (1980). *Springer Series on Medical Education/ Volume 1: Problem-Based Learning-An Approach to Medical Education.* New York, New York: Springer Publishing Company.

Bennett, R. E., & Ward, W. C. (1993). *Construction versus choice in cognitive measurement.* Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Bloom, B.S., Engelhart, M.D., Furst, E.J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain.* New York: David McKay.

Bloom, B. (1978) New view of learning: Implications for instruction and curriculum. *Educational Leadership, 35,* 562-568.

Bloom, B. S., & Broder, J.L. (1950). *Problem-solving processes of college students.* Chicago: The University of Chicago Press.

Bodner, G. M. (1991). A view from chemistry. In M. U. Smith (Ed.), *Towards a unified theory of problem solving: Views from the content domains* (pp. 21-33). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bordage, G., & Page, G. (1987). An alternative approach to PMPs: the "key features" concept. In I. R. Hart & R. M. Harden (Eds.), *Further developments in assessing clinical competence* (pp. 59-75). Montreal, Canada: Can-Heal Publications Inc.

Bordage, G., & Lemieux, M. (1991). Cognitive structures of experts and novices. *Academic Medicine, 66(9)*, September Supplement, S70-S72.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 1,* 55-81.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5,* 121-152.

Claessen, H. F. A., & Boshuizen, H. P. A. (1985). Recall of medical information by students and doctors. *Medical Education, 19* (1), 61-67.

Collins, A. (1990). Reformulating testing to measure learning and thinking. In N. Frederiksen, R. Glaser, A. Lesgold & M.G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 75-87). Hillsdale, NJ. Lawrence Erlbaum Associates.

Connolly, J. A., & Wantman, M. J. (1964). An exploration of oral reasoning processes in responding to objective test items. *Journal of Educational Measurement, 1,* 59-64.

Cox, K. (1978). How did you guess: What do MCQ's measure? *Medical Journal of Australia, 1,* 884-886.

Cronbach, L. J. (1971). Test validity. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Cutler, P. (1979) *Problem-solving in clinical medicine: From data to diagnosis*. Baltimore: Williams and Wilkins.

de Groot, A. D. (1965). *Thought and choice in chess.* Den Haag, The Netherlands: Mouton Publishers.

DuBois, P. H. (1970). *A History of Psychological Testing.* Boston: Allyn and Bacon, Inc.

Ebel, R. L. (1972). Essentials of educational measurement (2nd Ed.). Englewood Cliffs, NJ: Prentice-Hall.

Eichna, L. W. (1980). Medical-school education. *The New England Journal of Medicine, 303* (13), 727-753.

Elstein, A.S. (1993). Beyond multiple choice questions and essays:the need for a new way to assess clinical competence. *Academic Medicine, 68*(4), 244 - 249.

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). *Medical problem solving: An analysis of clinical reasoning.* Cambridge: Harvard University Press.

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1990). Medical problem solving: a ten year perspective. *Evaluation & The Health Professions. 13* (1), 5-36.

Embretson, S. (1983). Construct validity: construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179-197.

Ericsson, K. A. (1987). Theoretical implications from protocol analysis on testing and measurement. In R. Ronning, J. Glover, J.C. Conoley, & J. Witt (Eds.), *The influence of cognitive psychology on testing and measurement: The Buros-Nebraska Symposium on measurement and testing* (pp. 191-226). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ericsson, K.A., & Smith, J. (1991). Prospects and limits of the empirical study of expertise: an introduction. In K.A. Ericsson and J. Smith (Eds.), *Toward a general theory of expertise* (pp. 1-38). New York: Cambridge University Press.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215-251.

Evans, D.A., & Patel, V.L. (1989). *Cognitive science in medicine.* Cambridge, MA.:The MIT Press.

Farr, R., Pritchard, R., & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement, 27*(3), 209-226.

Feinberg, L. (1990). Multiple-choice and its critics. *The College Board Review* (No. 157), pp. 13-17, 30-31.

Frechtling, J. A. (1991).. Performance assessment:Moonstruck or the real thing?. *Educational Measurement:Issues and Practice, 10(4)*, 23-25.

Frederiksen, C. H. (1975). Representing logical and semantic structure of knowledge acquired from discourse. *Cognitive Psychology, 7,* 371-458.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39,* 193-202.

Garner, R. (1982). Verbal-report data on reading strategies. *Journal of Reading Behavior, XIV*, 159-167.

Garner, R. (1988). Verbal-report data on cognitive and metacognitive strategies. In C.E. Weinstein, E.T. Goetz, & P.A. Alexander (Eds.) *Learning and study strategies.* (pp. 63-76) San Diego, CA: Academic Press, Inc.

Giannini, G., & Engel, J. D. (1986). On the meaning of scores derived from patient management problems. *Evaluation & The Health Professions, 9,* 103-120.

Goodenough, F.L. (1949). *Mental testing.* New York, NY.: Rinehart.

Greeno, J.G. (1991). A view of mathematical problem solving in school. In M. U. Smith (Ed.), *Towards a unified theory of problem solving: Views from the content domains* (pp. 69-98). Hillsdale, NJ: Lawrence Erlbaum Associates.

Groen, G. J., & Patel, V. L. (1985). Medical problem-solving: some questionable assumptions.*Medical Education, 19* (2), 95-100.

Groen, G. J., & Patel, V. L. (1988). The relationship between comprehension and reasoning in medical education. In M. Chi, R. Glaser, & M. Farr (Eds.), *The nature of expertise* (pp. 287-310). Hillsdale, NJ: Erlbaum. .

Groen, G. J., & Patel, V. L. (1991). In M. U. Smith (Ed.), *Towards a unified theory of problem solving: Views from the content domains* (pp. 35 - 44).Hillsdale, NJ: Lawrence Erlbaum Associates.

Haladyna, T. M., & Downing, S. M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education. 2,* 37-50.

Haladyna, T. M., & Downing, S. M. (1989b). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education. 2,* 51-78.

Hambleton, R.K., & Murphy, E. (1992). A psychometric perspective on authentic measurement. *Applied Measurement in Education, 5*(1), 1-16.

Haney, W., & Scott, L. (1987). Talking with children about tests: An exploratory study of test item ambiguity. in R. O. Freedle & R. P. Duran (Eds.), *Cognitive and linguistic analyses of test performance* (pp. 298-368). Norwood, NJ: Ablex.

Harvill, L. M. (1985). Assessing the test-wiseness of health science students. *Evaluation & the Health Professions, 8,* 494-508.

Hoffman, B. (1962). *The tyranny of testing.* New York: Crowell Collier and Macmillan, Inc.

Holyoak, K.J. (1991). Symbolic connectionism:toward third-generation theories of expertise. In K.A. Ericsson and J. Smith (Eds.), *Toward a general theory of expertise* (pp. 301-335). New York. NY.: Cambridge University Press.

Kassirer, J.P., & Kopelman, R.I. (1989). Cognitive errors in diagnosis:Instantiation, classification, and consequences. *The American Journal of Medicine, 86,* 433-441.

Kaufman, D. R., & Patel, V. L. (1991). Problem solving in the clinical interview: A cognitive analysis of the performance of physicians, residents and students.*Teaching and Learning in Medicine, 3* (1), 6-14.

Kintsch, W., & van Dijk, T. (1978). Toward a model of text comprehension and production. *Psychological Review, 85* (5), 363-394.

Kropp, R. P. (1956). The relationship between process and correct item-response. *Journal of Educational Research, 49,* 385-388.

Lemieux, M., & Bordage, G. (1986). Structuralisme et pedagogie medicale: Etude comparative des strategies cognitives d'apprentis-medicins. *Recherches semiotiques, 6(2),* 143-179. (Published article translated into English by G.Bordage, no date).

Lemieux, M., & Bordage, G. (1992). Propositional versus structural semantic analyses of medical diagnostic thinking. *Cognitive Science, 16,* 185-204.

Lemieux, M., & Bordage, G. (1993). Comparing the core and the peel of the same fruit. *Cognitive Science, 17,* 143-147.

Lesgold, A., Rubinson, H., Feltovich, P., Glaser, R., Klopfer, D., & Wang, Y. (1988). Expertise in a complex skill: Diagnosing x-ray pictures. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise* (pp. 311-342). Hillsdale, NJ: Lawrence Erlbaum Associates.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3,* 635-694 (monograph Supp. 9).

McGuire, C. (1963). Research in the process approach to the construction and analysis of medical examinations. *National Council on Measurement in Education Yearbook, 20,* 7-16.

McGuire, C. (1987). Written methods for assessing clinical competence. In I.R. Hart & R.M. Harden (Eds.), Further developments in assessing clinical competence (pp. 46-58). Montreal, Canada:Can-Heal Publications.

Maguire, T., Skakun, E., & Harley, C. (1992). Setting standards for multiple-choice items in clinical reasoning. *Evaluation & The Health Professions, 15*(4), 434-452.

Maguire, T., Hattie, J., & Haig, B. (1993). *Construct validity and achievement assessment.* Presentation at the Cognition and Assessment Conference, Queen's University, Kingston, Ontario.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: National Council on Measurement in Education, American Council on Education, & Macmillan Publishing Company.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 335-366). New York: National Council on Measurement in Education, American Council on Education, & Macmillan Publishing Company.

Mislevy, R.J. (1993). A framework for studying differences between multiple-choice and free-response test items. In R.E. Bennett and W.C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 75-106). Hillsdale, NJ.: Lawrence Erlbaum Associates.

Morgan, M. K., & Irby, D. M. (1978). *Evaluating clinical competence in the health professions.* Saint Louis: The C. V. Mosby Company.

Muller, S. (1984). Physicians for the twenty-first century. Report of the project panel on the general professional education of the physician and college preparation for medicine. *Journal of Medical Education, 59,* Part 2, November.

Muzzin, L. J., Norman, G. R., Feightner, J. W., Tugwell, P., & Guyatt, G. (1983). Expertise in recall of protocols in two specialty areas. *Proceedings of the 22nd Annual Conference on Research in Medical Education,* (pp. 122-127). Washington, D. C.

Muzzin, L. J., Norman, G. R., Jacoby, L. L., Feightner, J. W., Tugwell, P., & Guyatt, G. H. (1982). Manifestations of of expertise in recall of clinical protocols. *Proceedings of the 21st Annual Conference on Research in Medical Education,* Washington, D. C.

Neufeld, V. R. (1985). Written examinations. In V. R. Neufeld & G. R. Norman (Eds.), *Assessing clinical competence* (pp. 94-118). New York, NY: Springer Publishing Company.

Newell, A., & Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice Hall.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231-259.

Norman, G. R., Brooks, L. R., & Allen, S. W. (1989). Recall by expert medical practitioners and novices as a record of processing attention. *Journal of Experimental Psychology: Learning Memory Cognition, 13,* 1166-1174.

Norman, G. R., Brooks, L. R., Allen, S. W., & Rosenthal, D. (1989). The development of expertise in Dermatology. *Archives of Dermatology, 125,*1063-1068.

Norman, G. R., Jacoby, L.L., Feightner, J. W., & Campbell, E.J. M. (1979). Clinical experience and the structure of memory. *Proceedings of the 18th Annual Conference on Research in Medical Education,* Washington, D. C.

Norman, G. R., Tugwell, P., Feightner, J. W., Muzzin, L. J., & Jacoby, L. L. (1985). Knowledge and clinical problem-solving. *Medical Education, 19* (5), 344-356.

Norris, S. P. (1989a). Verbal reports of thinking and multiple-choice critical thinking test design (Technical Report No. 447). Champaign: University of Illinois at Urbana-Champaign, Center for the Study of Reading.

Norris, S. P. (1989b). Can we test validly for critical thinking? *Educational Researcher, 18,* 21-26.

Norris, S. P. (1990). Effect of eliciting verbal reports of thinking on critical thinking test performance. *Journal of Educational Measurement, 27,* 41-58.

Olson, J. R., & Biolsi, K. J. (1991). Techniques for representing expert knowledge. In K.A. Ericsson and J. Smith (Eds.), *Toward a general* *theory of expertise* (pp. 240-285). New York. NY.: Cambridge University Press.

Osterlind, S. J. (1989). *Constructing test items.* Boston, MA: Kluwer Academic Publishers.

Patel, V. L., Evans, D. A., & Groen, G. J. (1989). Biomedical knowledge and clinical reasoning. In D. A. Evans & V. L. Patel (Eds.), *Cognitive science in medicine:biomedical modeling.* (pp. 53-112). Cambridge, Massachusetts: MIT Press.

Patel, V. L., Evans, D. A., & Kaufman, D. R. (1990). Reasoning strategies and the use of biomedical knowledge by medical students. *Medical Education, 24,* 129-136.

Patel, V. L., & Groen, G. J. (1986). Knowledge based solutions strategies in medical reasoning. *Cognitive Science, 10*, 91-116.

Patel, V. L. Groen, G. J. & Arocha, J. F. (1990). Medical expertise as a function of task difficulty. *Memory & Cognition, 18* (4), 394-406.

Patel, V. L., Groen, G. J., & Frederiksen, C. H. (1986). Differences between medical students and doctors in memory for clinical cases.*Medical Education, 20* (1), 3-9.

Patel, V. L. Groen, G. J., & Norman, G. R. (1991). Effects of conventional and problem-based medical curricula on problem solving. *Academic Medicine, 66* (7), 380-389.

Patel, V. L., Groen, G. J., & Scott, H. M. (1988). Biomedical knowledge in explanations of clinical problems by medical students. *Medical Education, 22* (5), 398-406.

Patel, V.L, & Groen, G.J. (1991). The general and specific nature of medical expertise: a critical look. In K.A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise* (pp. 93-125). New York. NY.: Cambridge University Press.

Patel, V.L., & Groen, G.J. (1993). Comparing apples and oranges:Some dangers in confusing frameworks with theories. *Cognitive Science, 17*, 135-141.

Perkins, D. N., & Salomon, G. (1989). Are cognitive skills context-bound. *Educational Researcher, 18,* 16-25.

Phillips, L. M. (1989). Developing and validating assessments of inference ability in reading comprehension (Technical Report No. 452). Champaign: University of Illinois at Urbana-Champaign, Center for the Study of Reading.

Ramsden, P., Whelan, G., & Cooper, D. (1989). Some phenomena of medical students' diagnostic problem-solving. *Medical Education, 23,* 108-117.

Ridderikhoff, J. (1989). *Methods in medicine: A descriptive study of physicians' behavior,* Boston: Kluwer Academic Publishers.

Ridderikhoff, J. (1991). Medical problem-solving:An exploration of strategies. *Medical Education, 25*, 196-207.

Rogers, W. T., & Bateson, D.J. (1991). The influence of test-wiseness on performance of high school seniors on school leaving examinations. *Applied Measurement in Education, 4*(2), 159 - 183.

Royer, J.M., Cisero, C.A., & Carlo, M.S. (1993). Techniques and procedures for assessing cognitive skills. *Review of Educational Research, 63*, 201-243.

Sarnacki, R. E. (1981). The effects of test-wiseness in medical education. *Evaluation & The Health Professions, 4,* 207-221.

Schmidt, H.G., Hobus, P.P.M., Patel, V.L., & Boshuizen, H.P.A. (1987). Contextual factors in the activation of first hypotheses:Expert-novice differences. Presentation at the Annual Meeting of the American Educational Research Association, Washington, DC.

Schmidt, H. G., Norman, G. R., & Boshuizen, H. P. A. (1990). A cognitive perspective on medical expertise: Theory and implications. *Academic Medicine, 65* (10), 611-621.

Schultz, K., & Lochhead, J. (1991). A view from physics. In M. U. Smith (Ed.), *Towards a unified theory of problem solving: Views from the content domains* (pp. 99-114). Hillsdale, NJ: Lawrence Erlbaum Associates.

Schuman, H. (1966). The random probe: A technique for evaluating the validity of closed questions. *American Sociological Review, 31,* 218-222.

Shepard, L.A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of Research in Education* (pp. 405-450). Washington, DC: American Educational Research Association.

Shimberg, B. (1985) Overview of professional and occupational licensing. In J.C. Fortune and Associates (Eds.), *Understanding testing in occupational licensing.* San Francisco, CA: Jossey-Bass.

Siegler, R. S. (1986). Unities across domains in children's strategy choices. In M. Perlmutter (Ed.), *Minnesota symposium on child development, Vol. 19* (pp.1-48). Hillsdale, NJ: Erlbaum.

Siegler, R. S. (1989). Strategy diversity and cognitive assessment. *Educational Researcher, 18*, 15-20.

Siegler, R. S., & McGilly, K. (1989). Strategy choices in children's time-telling. In I. Levin & D. Zakay (Eds.), *Time and human cognition: A life-span perspective* (pp.185-218). New York: North-Holland.

Siegler, R. S., & Robinson, M. (1982). The development of numerical understanding. In H. Reese & L. P. Lipsett (Eds.), *Advances in child development and behavior. Vol. 16,* New York, NY: Academic Press.

Siegler, R. S., & Shrager, J. (1984). Strategy choices in addition and subtraction: How do children know what to do? In C. Sophian (Ed.), *Origins of cognitive skills.* Hillsdale, NJ: Erlbaum.

Skakun, E. N. (1982). Relating the components of the clinical reasoning process to the construction of patient management problems. *Proceedings of the sixth annual symposium on computer applications in medical care* (pp. 732-736). Washington, D. C.: Computer Society Press.

Smith, M. U. (1991). A view from biology. In M. U. Smith (Ed.), *Towards a unified theory of problem solving: Views from the content domains* (pp. 1- 19). Hillsdale, NJ: Lawrence Erlbaum Associates.

Snow, R.E. (1993). Construct validity and constructed-response tests. In R.E. Bennett & W.C. Ward (Eds.), *Construction versus choice* *in cognitive measurement* (pp. 45-60). Hillsdale, NJ.: Lawrence Erlbaum Associates.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263-331) New York: National Council on Measurement in Education, American Council on Education, & Macmillan Publishing Company.

Stoker, H. W., & Kropp, R. P. (1964). Measurement of cognitive processes. *Journal of Educational Measurement, 1,* 39-42.

Swanson, D. B., & Stillman, P. L. (1990). Use of standardized patients for teaching and assessing clinical skills. *Evaluation & The Health Professions, 13* (1), 79-103.

The Medical Letter on Drugs and Therapeutics (M. Abramowicz (Ed.). (1991). Drugs for rheumatoid arthritis. New Rochelle, NY: The Medical Letter Inc. (Vol 33, p. 65-70).

Tittle, C.K., Hecht, D., & Moore, P. (1993). Assessment theory and research for classrooms:From taxonomies to constructing meaning in context. *Educational Measurement:Issues and Practice, 12*(4), 13-19.

Traub, R.E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R.E. Bennett & W.C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 29-44). Hillsdale, NJ.: Lawrence Erlbaum Associates

Wainer, H., & Braun, H. I. (1988). *Test validity*. Hillsdale, NJ: Erlbaum.

Ward, W. C. (1985). Measurement research that will change test design for the future. In E.E. Freeman (Ed.), *The redesign of testing for the 21st century* (pp. 25-34). Princeton, NJ: Educational Testing Service.

Wesman, A. G. (1971). Writing the test item. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 81-129). Washington, DC: American Council on Education.

Appendix I

Interviewing Model for Multiple Choice Question Strategies

Step One

1.1 Inform examinee of purpose of interview: to find out how the examinee approaches multiple choice questions, to indicate the process being used and to find out what the examinee is thinking while choosing answers to questions on the test.

1.2 Inform examinee of his or her role: to select the best answer, to respond as completely as possible to how the examinee is approaching the multiple choice questions and arriving at the best answer.

Step Two

2.1 Interviewer says to examinee:
"As you do each question tell me all you can about what you are doing mentally as you attempt to answer each question and your thinking (i.e., justification) while you are selecting your answers."

2.2 Interviewer can interrupt examinee's narrative only to:

* probe for ambiguous reference by saying:
"Could you tell me what you mean by....?
Example: When doing Item 3 the examinee says: "This option is correct."
Probe immediately: "Could you tell me by number which option is correct?"

* probe for obvious reading mistakes by saying:
Did you read....?" (Do not endorse answers)
Example: When doing Item 1 the examinee reads: "Mrs. X is in hospital for "Mrs. X is not in hospital." Probe immediately: "Did you read 'Mrs. X is in hospital?"

* probe for part of multiple choice question that the student is processing.

Example: Student narrates that he or she is reading option 3, but then inquires about data in stem, probe by asking "Could you tell me what part of the item you were reading prior to your question?"

* probe for verbalizations

Example: Student says correct answer without verbalizing how question was processed, probe by asking "Could you start again and this time tell me where you started and which part of the question you read first?"

Example: Student gives correct answer, pauses and asks "do I go on to the next item?" You would reply "Before you go on to the next item, could you tell me what you were thinking when you were answering item 4."

2.3 Interviewer can respond to examinee's questions only as follows:

* If examinee probes for fact, say: "you can only go by what is written."
Example: When doing Item 1 the examinee asks: "How long has Mrs. X been in hospital?" Answer only: "You can only go by what is written."

* If the examinee probes for reason, say: "You can decide only according to what is said and what you know."
Example: When doing Item 1 the examinee asks: "Why was a WBC ordered?" Answer only: "You can decide only according to what is written and what you know."

2.4 General cautions:

* Do not begin to speak immediately after the examinee stops; give the examinee a few seconds to continue.

* Do not cut off examinee's reasoning by in any way signalling that enough has been said, even though the examinee might seek such signals.

* Do not endorse or criticize the examinee's approach to items, his her choice of correct answer, and justification for correct answer.

Step Three

3.1 Inform examinee when to proceed to the next item. When examinee has given the answer, verbalized approach to item, and verbalized thinking processes in responding to item, interviewer says "You can now proceed to item 2."

Appendix II

Introduction to Examinees

THE USE OF STRATEGIES IN RESPONDING TO MULTIPLE CHOICE ITEMS

Background

Multiple choice (MC) items continue to be used on many examinations. Many of us remember multiple choice examinations administered by our teachers as we progressed through junior high and senior high school. In the Faculty of Medicine, MC items can be found on course examinations. The Comprehensive Examinations administered at the conclusion of each Phase are exclusively MC.

Although MC items have been around for a long time, very little is known about what transpires when examinees encounter multiple choice items. What we are attempting to do in this pilot study is identify the **strategies** and the **thinking process** medical students use when responding to MC items. To accomplish this we will present to you a set of one best or correct answer items. These items consist of a statement, also known as a stem, and 5 alternatives or plausible answers. Of these 5 alternatives, only one is considered to be the best answer or correct answer. If you would like to refresh your memory about this type of item, examples appear on page 7.

As you read and answer each question, you will be asked to "think aloud" or verbalize what you are doing, what you are reading, and what you are thinking. To familiarize you with the "think aloud" aspect, scenarios of several different "think aloud" protocols for an an item will be given by the interviewer.

Before going on, we would like to thank you for cooperating and participating in this pilot study. In order for us to identify the strategies and the processes used by different students the entire interview will be audio-taped. You are assured that the interview material will be held in confidentiality and that it will not be made public.

Instructions

1. Purpose of Interview. We are interested in how you approach multiple choice items. For each item, we want to know which parts of the item you are reading and what you are thinking as you read the item. We want to know how you arrived at the correct or best answer.

2. Role of the Student. Your task is to tell me the best or correct answer to each multiple choice question presented to you. For each item you are to tell me what part or parts of the item you are reading and what you are thinking of as you are going through the item. You will also have to justify or explain why the answer you have selected is correct or the best. If you have any questions as you proceed through each item, do not hesitate to ask them, however, the interviewer will not provide any more information than presented by each item. Tell me when you are finished with the item and ready for the next one.

Since it is very likely that you have not participated in an exercise of this type before, I will provide you with several different "thinking aloud" scenarios or protocols for a sample item. However, before I do that, do you have any questions thus far?

3. Sample "Thinking Aloud" Scenarios. The interviewer will present several examples of think aloud protocols for several items.

Appendix III

Sample Think Aloud Protocols

Scenario - Sample "Think Aloud" Protocols

A 40-year-old patient was admitted to the hospital complaining of recent onset of dyspnea, right-sided chest pain aggravated by deep inspiration and hemoptysis.

The likely diagnosis is:

1. Myocardial infarction.
2. Pericarditis.
3. Pulmonary embolism with infarction.
4. Mitral stenosis.
5. Pulmonary tuberculosis.

EXAMPLE 1

This is an A-type item. Usually what I do with these types of items, I cover the options and read the top part first. I am now reading the stem. So this is a 40-year old patient. I don't know whether the patient is male or female and you tell me that I can only go by what is given in the item. This patient has chest pain and shortness of breath. The presentation suggests an acute condition because it says recent onset, although we don't know how recent. The pain is pleuritic suggesting an irritation of the parietal pleura. There is an expectoration of blood. We need a diagnosis. Well the hemoptysis suggests infarction of lung tissue. Now let me look at the options. The first one of myocardial infarct is not very likely because the pain is usually central but in this patient it is right-sided. Now, the pain, the pain in myocardial infarct is not aggravated by inspiration and it's not associated with hemoptysis, so I don't think it's option 1. Option 2, that is a possibility, pericardial pain could be aggravated by breathing but the pain is usually central, it may be pleuritic but it is not associated with blood stained sputum. Oh, another thing, usually there is no dyspnea, so if I put all this together, option 2 doesn't look right either. Number 3, pulmonary embolism with infarction, there it is, it looks like the best answer on the basis of the history but let me look at the other options. Mitral stenosis - could present with hemoptysis and dyspnea but not with pleuritic pain and in the patient there is pleuritic pain. No, not option 4. Pulmonary TB, it could be, but the onset of TB is not acute, it is more of an insidious onset and what we have here is a recent one, so TB is not very likely. Yes, I would say that 3, pulmonary embolism with infarction is most likely.

EXAMPLE 2

I first read the paragraph, the introduction to the question. As I read this I high-light the important things like age and complaints. I note the symptoms - shortness of breath, chest pain, and the character of the pain. This looks like a picture of pulmonary embolism, my differential diagnosis is pulmonary embolism. It asks me for a diagnosis. So now I will look at the choices. I am going to search for pulmonary embolism. OK, MI, that is not embolism and MI does not fit with the symptoms, so it's not MI. Choice 2 does not fit. It does not match my differential, does not fit the symptoms, 2 is wrong. Three, pulmonary embolism with infarction, that matches my differential diagnosis, that is the correct answer. I'll just do a quick check of 4 and 5 to make sure I have not skipped something. Mitral stenosis is incorrect so is 5 TB. Three is correct.

EXAMPLE 3

I'll start at the top. This is a 40 year old person who has had a productive life, has been well and asymptomatic until recently. There is shortness of breath, pain aggravated on inspiration and hemoptysis. They want a diagnosis. It is pulmonary embolism. I'll just scan the answers, I see pulmonary embolism, 3 is correct, the rest are wrong. But you'll never see a patient walk off the street like this, this patient would have to be lying down for a long time.

EXAMPLE 4

I am going to read the stem. I have the age and the setting. It is an acute onset and involves the lung or the heart, most likely the lung but I can not exclude the heart. That's the stem. The most likely diagnosis is:.... myocardial infarct, sudden onset, pain but not aggravated by deep inspiration, MI is not likely. Pericarditis is possible, pulmonary embolism is possible, mitral stenosis - NO, and 5 is obviously a distractor. So it's 2 or 3. I would go with pericarditis on the basis of age, pain, and the character of the pain. NO!, I can't do that! There is no hemoptysis in pericarditis. It's 3 - Pulmonary embolism.

Appendix IV

Practice and Experimental Items

PRACTICE ITEMS for study

Item 1

A 35-year-old woman presents with signs, symptoms and laboratory tests that indicate mild hyperthyroidism. On examination, the thyroid gland is normal in size and texture, but very tender to palpation. The most appropriate test to perform to aid in the diagnosis is:

1. Fine needle aspiration therapy biopsy of the thyroid.
2. 131I thyroid uptake.
3. Ultrasound of the thyroid.
4. Thyroid antibodies.
5. Serum thyroglobulin.

Item 2

During the past winter, a 48-year-old housewife had two attacks of "pneumonia" with hemoptysis. She recovered very slowly and now notices constantly irregular pulse and dyspnea on effort.

Her past history would probably reveal:

1. A murmur detected in infancy and a family history of congenital heart disease.
2. Prolonged and excessive use of alcohol.
3. Syphilis detected in the secondary stage and inadequately treated with penicillin.
4. A heart murmur noted from birth.
5. A history of rheumatic fever as a child.

EXPERIMENTAL ITEMS for study

Item Id. 582 Item 1

One week after an anterior myocardial infarction a 55-year-old man complains of severe pain in the left leg. The leg is cool, pale and pulseless.

The most likely diagnosis is:

1. Deep venous thrombosis.
2. Ruptured left iliac aneurysm.
* 3. Arterial embolism.
4. A-V fistula.
5. Arterial thrombosis.

Item Id. 1085 Item 2

The most likely physical finding to be noted in a patient during an attack of angina on effort is a:

* 1. Fourth heart sound.
2. Second sound of diminished intensity.
3. Mid-systolic murmur at the cardiac apex.
4. Transient aortic ejection click.
5. Third heart sound.

Item Id. 15 Item 3

A 65-year-old male who has a history of heavy cigarette consumption for many years presents with pulmonary emphysema and intercurrent acute bronchitis. In addition to the signs of chronic lung disease, he has massive peripheral edema, enlarged pulsating liver, elevated jugular venous pressure, and large V waves in the neck.

His x-ray may be expected to show:

1. Acute congestion of the pulmonary veins.
2. Enlarged left ventricle with increased aeration of the lungs.
3. Straightening of the right border of the heart by right ventricular enlargement.
4. Bi-ventricular enlargement and normal pulmonary vessels with flattening of the diaphragm.

* 5. "Pruning" of smaller branches of the pulmonary artery and flattening of the diaphragm.

Item Id. 11 Item 4

The most frequent site of subluxation of the vertebrae in patients with rheumatoid arthritis is:

* 1. C1-2

2. C5-6
3. L4-5
4. L5-S1
5. L5-S2.

Vignette 909 Item 5
Related Items 909 and 911
Item Id. 909

A 22-year-old male has had intermittent low back pain for 4 years. It does not seem to be related to anything in particular and usually lasts 2-3 days, and he is now concerned because his right eye is red and painful and the vision is decreasing.

Which of the following is your probable diagnosis of his systemic disease?

1. Rheumatoid arthritis.
* 2. Ankylosing spondylitis.
3. Tuberculosis of the spine.
4. Sarcoidosis.
5. Reiter's syndrome.

Item Id. 911 Item 6

A 22-year-old male has had intermittent low back pain for 4 years. It does not seem to be related to anything in particular and usually lasts 2-3 days, and he is now concerned because his right eye is red and painful and the vision is decreasing.

Your most important diagnostic test to establish the diagnosis of this disease is:

1. X-ray of the chest.
2. HLA typing.
* 3. X-ray of the spine.
4. Chlamydial antibody titres.
5. Scalene node biopsy.

Item Id. 612 Item 7

A 70-year-old man has had a smoker's cough for many years. Two weeks ago his voice became weak and hoarse, and one week ago he began to cough up 50 to 100 g of purulent sputum daily. A chest x-ray reveals a cavity containing an air-fluid level in the left upper lobe. He is afebrile.

Which of the following would you advise?

1. Surgery with removal of the affected lobe.

* 2. Bronchoscopy.

3. Sputum for acid fast bacilli culture daily for three days.
4. Intensive antibiotic therapy.
5. Postural drainage four times daily.

Item Id. 32 Item 8

Which of the following drugs is generally NOT useful in improving the long term outlook in patients with rheumatoid arthritis?

1. Penicillamine.
2. Gold.

* 3. Prednisone.

4. Chloroquine.
5. Methotrexate.

Item Id. 568 Item 9

In patients with non-penetrating cardiac trauma, the most common cardiac injury is laceration or rupture of the:

1. AV valve.
2. Semilunar valve.
3. Coronary artery.
4. Left ventricle.

* 5. Papillary muscle.

Item Id. 1076 Item 10

Which of the following series of heart sounds is in the correct sequence? (O.S. = Opening Snap)

1. 1st, 2nd, 3rd, 4th, O.S.
2. 1st, A2P2, 3rd, 4th, O.S.
3. 3rd, 1st, A2P2, O.S., 4th.

* 4. 4th, 1st, A2P2, O.S., 3rd.

5. 4th, 1st, P2A2, O.S., 3rd.

Item Id. 1950 Item 11

At present, the agent of choice for MRSA (Methicillin resistant Staphylococcus aureus) is:

1. Cloxacillin.

* 2. Vancomycin.

3. Ceftazidime.
4. Erythromycin.
5. Cephalothin.

Item Id. 385 Item 12

A 75-year-old man gives a long history of intermittent diarrhea, some constipation, and lower abdominal pain. Recently the pain has become more severe and has been associated with urinary frequency and burning. On physical examination the temperature is 39°C and a tender mass is palpable in the left lower quadrant of the abdomen.

A complication likely to develop in this patient is:

* 1. Vesico-colic fistula.

2. Massive hemorrhage.
3. Subphrenic abscess.
4. Jaundice.
5. Urinary obstruction.

Item Id. 381 Item 13

A 72-year-old man presents himself at your office complaining of weakness, easy fatigue and mild abdominal discomfort. The only finding on physical examination is pallor of skin and mucous membranes. Blood hemoglobin is 60 g/L (6.0 g/dL). A stained smear of the peripheral blood shows numerous microcytes. Rectal examination shows only an enlarged prostate.

Your first action would be to:

1. Examine the bone marrow for iron stores.
2. Determine the serum ferritin level.
3. * Examine the stool for gross and occult blood.
4. Determine the serum iron level.
5. Administer blood transfusion.

Item Id. 733 Item 14

A 56-year-old man presents to his doctor with a month history of intermittent right facial pain. On examination he is found to have a diminished corneal reflex and slight hearing defect on the right.

The diagnosis is:

1. Right cerebral tumor.
2. Trigeminal neuralgia.
3. Otitis media.
4. * Acoustic neuroma.
5. Multiple sclerosis.

Item Id. 347 Item 15

A previously healthy 27-year-old female is suddenly seized with pleuritic pain in the left chest and shortness of breath.

The most likely cause is:

1. Mycoplasma pneumonia.
2. * Spontaneous pneumothorax.
3. Pulmonary embolism.
4. Acute pericarditis.
5. Epidemic pleurodynia.

Item Id. 1033 Item 16

A 24-year-old nurse who previously had a negative tuberculin skin test is found to have a positive 5 T.U. Mantoux test three months after exposure to a patient with active tuberculosis. She is asymptomatic and has a normal chest x-ray. Sputum is negative for Mycobacterium tuberculosis both on smear and culture. The best management would consist of:

1. Repeat chest x-ray at yearly intervals.
2. * Administration of isoniazid for 12 months.
3. Repeat Mantoux test with 1 T.U. in 4 weeks.
4. Administration of isoniazid and ethambutol for 12 months.
5. Repeat chest x-ray every three months and start treatment if radiologic evidence of tuberculosis appears.

Item Id. 267 Item 17

The investigation of bilateral gynecomastia in a 37-year-old male should involve determining all of the following, EXCEPT:

1. The serum testosterone level.
2. A history of marijuana usage.
3. The serum chorionic gonadotropin level (hCG).
4. The patient's thyroid status.
5. * Ultrasound of the abdomen.

Item Id. 1101 Item 18

A patient on a mechanical ventilator has the following blood gases: pO2 70 mm Hg, pCO2 40 mm Hg, [H+] 58 nmol/L (pH 7.24), oxygen saturation 91%.

Which one of the following therapeutic maneuvers would you favor?

1. Addition of one length of dead space to ventilator circuit.
2. Administration of oxygen.
3. Administration of intravenous glucose in normal saline.
4. * Administration of intravenous sodium bicarbonate.
5. Administration of a CO2 - O2 mixture.

Item Id. 1572 Item 19

A 24-year-old airline flight attendant complains of feeling tired and losing weight in spite of a good appetite. For the past year she has noticed voluminous, pale, foul-smelling stools. She recalls being told of having bowel difficulty in early childhood and of being fed a diet consisting largely of bananas.

Radiological examination discloses an abnormal small bowel follow through. Biochemical analysis of the stool shows an increased amount of fat. The blood picture shows anemia.

Which of the following diets would you select for this patient?

* 1. Gluten free.
2. Lactose free.
3. Low fat.
4. Low residue.
5. High residue.

Item Id. 1034 Item 20

The most definitive diagnostic test for pulmonary embolism with or without infarction is:

1. Perfusion lung scan.
2. Ventilation-perfusion lung scan.
3. A decreased arterial pCO2.
4. An increased alveolar-arterial oxygen difference.
* 5. Pulmonary angiogram.

Item Id. 409 Item 21

A 63-year-old man first noted difficulty swallowing about six months ago. He states that initially bread and meat became "stuck" but could be washed down with liquids. For the past month or so, he has subsisted only on liquids. Over the past several months he has lost 11 kg.

What diagnosis would you consider most likely?

1. Peptic esophageal stricture.
2. Achalasia.
3. Esophageal web.
4. Diffuse esophageal spasm.

* 5. Carcinoma of the esophagus.

Item Id. 1980 Item 22

A positive direct Coombs' test implies:

1. That the patient has increased antibodies (immunoglobulins) in the serum.
2. That the patient has antibodies in the serum that are directed against red cell antigens.

* 3. That the patient's red cells are coated with antibody.

4. That decreased red cell survival is present (i.e. hemolysis).
5. All of the above.

Item Id. 395 Item 23

In trigeminal neuralgia, which of the following is the treatment of choice in the early stages?

* 1. Carbamazepine (Tegretol).

2. Phenytoin (Dilantin).
3. Acetylsalicylic acid.
4. Pethidine hydrochloride (Demerol).
5. Alcohol injection of trigeminal ganglion.

Item Id. 1947 Item 24

Within a few days or a few weeks of onset, an upper motor neuron lesion is often characterized by:

1. Absent deep tendon reflexes on the affected side.
2. The presence of fasciculation in the affected limb.
3. Wasting of muscles.
- * 4. The presence of clonus in the affected limb.
5. Flapping tremor in the affected limb.

Item Id. 206 Item 25

A diagnostically helpful ophthalmic finding in lupus erythematosus is:

* 1. Cytoid bodies.
2. Microaneurysms.
3. Roth spots.
4. Macular degeneration.
5. Nystagmus.

Item Id. 1098 Item 26

Which of the following glomerulopathies with renal insufficiency is most likely to benefit from corticosteroids or azathioprine?

1. Goodpasture's syndrome.
* 2. Lupus nephritis.
3. Pre-eclampsia.
4. Amyloidosis.
5. Scleroderma.

Item Id. 332 Item 27

A 28-year-old keypunch operator has a history of having had pneumonia four times in the past twenty years. She has had a cough "all her life" which is worse in winter.

Physical examination reveals dullness, diminished breath sounds and numerous crepitations below T3 bilaterally. Her fingers are clubbed.

She probably has:

1. Hypogammaglobulinemia.
2. Congenital heart disease.
3. Bronchiolitis obliterans.

* 4. Bronchiectasis.

5. Cystic fibrosis.

Item Id. 773 Item 28

The most appropriate topical agent in the treatment of severe dermatitis is:

1. Calamine lotion.
2. Vioform-hydrocortisone combination.

* 3. Corticosteroid.

4. Antibiotic-corticosteroid combination.
5. Colloidal oatmeal bath.

Item Id. 392 Item 29

Generalized pruritus is most commonly due to:

1. Hodgkin's disease.
2. Polycythemia.
3. Hyperthyroidism.
4. Chronic renal failure.

* 5. Xerosis (dry skin).

Item Id. 771 Item 30

The irritable bowel syndrome in adults ("irritable colon") is a diagnosis of exclusion. However, when this diagnosis is finally made you should:

1. Tell the patient the symptoms are always due to emotional stress.
2. Tell the patient to routinely take tranquilizers when symptoms flare.
3. Tell the patient to return for a complete reevaluation (x-rays, blood work, etc.) in three months.
4. * Counsel the patient and prescribe metamucil and bran.
5. Counsel the patient and prescribe Lomotil and Kaopectate.

Appendix V

Student Questionnaire

QUESTIONNAIRE

Now that you have completed the exercise related to the items, we would like your opinion about several aspects of the study and the test items. Each of these aspects is described by sets of adjective pairs or sets of descriptions. For each pair of adjectives or set of descriptions, specify your opinion using the four-point scale. In the following example, dealing with INSTRUCTIONS and the adjective pair **good - bad**, you would record your opinion in the following manner if you felt that the instructions were good.

INSTRUCTIONS
good _**X**_ : ___ : ___ : ___ bad

Complete the following:

PURPOSE OF STUDY

1. explained clearly ___ : ___ : ___ : ___ lacked clarity
2. informative ___ : ___ : ___ : ___ uninformative

INSTRUCTIONS

3. clear ___ : ___ : ___ : ___ ambiguous
4. complex ___ : ___ : ___ : ___ simple
5. easy to follow ___ : ___ : ___ : ___ difficult to follow

THINK ALOUD EXAMPLES

6. useless ___ : ___ : ___ : ___ helpful

7. influenced my thinking ___ : ___ : ___ : ___ did not influence my thinking

8. influenced my responses ___ : ___ : ___ : ___ did not influence my responses

WHEN THINKING ABOUT THE TEST ITEMS

9.	probing by the researcher interfered with my reasoning	___ : ___ : ___ : ___	probing by the researcher did NOT interfere with my reasoning
10.	my thinking and reasoning were typical of how I approach exams	___ : ___ : ___ : ___	my thinking and reasoning were NOT typical of how I approach exams

WHEN THINKING ABOUT THE TEST ITEMS

11.	I used strategies that I typically use on other exams	___ : ___ : ___ : ___	I did NOT use strategies that I typically use on exams
12.	I mirrored my thinking on the examples	___ : ___ : ___ : ___	I did NOT mirror my thinking on the examples provided
13.	**This exam was:**		
	difficult	___ : ___ : ___ : ___	easy
	testing essentials	___ : ___ : ___ : ___	testing obscurity
	out-dated	___ : ___ : ___ : ___	current
	fair	___ : ___ : ___ : ___	unfair
	like other exams	___ : ___ : ___ : ___	unlike other exams

14. **The examination content represented:**

course or rotation objectives ___ : ___ : ___ : ___ no course or rotation objectives

material taught in classes ___ : ___ : ___ : ___ material not taught in classes

material presented on the ward ___ : ___ : ___ : ___ material not presented on the ward

15. **The test items were:**

clear ___ : ___ : ___ : ___ ambiguous

incomplete ___ : ___ : ___ : ___ complete

poorly designed ___ : ___ : ___ : ___ properly designed

16. **Comments:**

____

____

____

Appendix VI

Examples of Multiple Choice Items

EXAMPLES OF MULTIPLE CHOICE ITEMS

The Item consists of a stem and 5 options of which only ONE is the best or correct answer. Here are some examples:

1. In severe aortic insufficiency, which of the following may be anticipated on physical examination?
    1. A slowed arterial pulse upstroke.
    2. An early decrescendo diastolic murmur at the left sternal border.
    3. An opening snap and loud first heart sound.
    4. A widely split second sound which does not vary with respiration.
    5. Clubbing of the fingers.

Response is 2.

2. Which of the following central nervous system cell populations is the most rapidly affected by ischemia?
    1. Axis cylinders.
    2. Astrocytes.
    3. Microglia.
    4. Oligodendroglia.
    5. Nerve cell bodies.

Response is 5.

3. A 70-year-old man becomes ill while in hospital for investigation of chronic urinary retention with infection. He is found to be cold, disoriented, and dyspneic. Blood pressure is 100/80; pulse is 110/min., central venous pressure 5cm H_2O, hemoglobin 12g/100mL, WBC 11 000/mm^3. He has not secreted any urine for the past two hours. The cause of his deterioration is most likely because of:
    1. Acute urinary retention.
    2. Acute uremia.
    3. Dehydration.
    4. Septic Shock
    5. Congestive cardiac failure.

Response is 4.

Appendix VII

Frequency Counts - Activities by Item

Number of moves - Items 1,2,3,4,5,7, 8, and 9

ACTIVITY	ITEMS							
	1	2	3	4	5	7	8	9
A. ITEM RELATED								
1. Reads item	4	15	1	3	-	9	20	14
2. Read and search	8	39	33	27	14	37	39	31
3. Generate - trigger	7	-	-	1	3	2	-	9
4. Generate - no reason	5	1	-	10	7	-	-	-
5. Generate - reason	20	-	7	2	15	1	-	-
B. ALTERNATIVES								
1. No comment	12	32	21	19	9	12	7	12
2. No	11	29	7	6	8	7	1	19
3. Do not know	3	4	8	6	3	-	1	7
4. Scans	4	-	-	-	-	-	-	-
5. Guesses	-	4	-	4	-	-	2	1
6. Ignores	2	-	1	65	13	1	4	-
7. Delimiting	-	-	-	-	16	-	1	-
8. Predisposing	2	1	-	-	1	-	-	-
9. Oppositions	4	2	5	-	2	2	140	4
10. Similarities	-	-	-	-	-	2	-	11
11. Category	6	20	1	35	1	5	-	13
12. Priority	-	-	-	-	-	84	-	-
13. Likelihood	39	8	9	18	9	5	2	52
14. Answer	10	-	-	-	10	-	-	-
15. Description	24	17	47	6	37	6	-	8
16. Association	43	47	61	5	47	36	-	34
C. PROBLEM SOLVING								
1. Restates	32	23	30	5	14	17	8	8
2. Redescribes	26	19	32	8	10	25	-	16
3. Generates answer-stem	32	1	7	13	25	3	-	9
4. Activates hypothesis	-	-	39	-	-	33	-	-

Number of moves (Continued) - Items 10 - 16

ACTIVITY	ITEMS						
	10	11	12	13	14	15	16
A. ITEM RELATED							
1. Reads item	3	7	6	7	7	7	6
2. Read and search	38	20	36	36	19	19	21
3. Generate - trigger	-	-	1	2	13	3	7
4. Generate - no reason	2	17	-	-	1	4	2
5. Generate - reason	3	3	2	5	14	10	
B. ALTERNATIVES							
1. No comment	11	18	11	17	10	22	9
2. No	5	11	20	14	9	17	13
3. Do not know	-	1	5	1	1	25	5
4. Scans	3	11	-	4	-	2	15
5. Guesses	4	-	-	-	-	-	-
6. Ignores	7	73	5	3	6	1	1
7. Delimiting	-	-	-	3	10	5	-
8. Predisposing	-	-	-	-	-	4	1
9. Oppositions	-	2	-	-	-	35	-
10. Similarities	12	11	-	6	-	-	-
11. Category	-	12	-	3	1	-	13
12. Priority	-	9	-	76	-	-	46
13. Likelihood	2	7	39	6	20	12	13
14. Answer	14	1	-	2	-	-	8
15. Description	106	-	22	6	51	22	4
16. Association	-	4	58	19	52	15	32
C. PROBLEM SOLVING							
1. Restates	4	-	14	12	5	3	4
2. Redescribes	18	-	33	35	34	3	36
3. Generates answer-stem	2	20	4	4	19	21	19
4. Activates hypothesis	-	-	38	38	-	-	-

Number of moves (Continued) - Items 17 - 23

ACTIVITY	ITEMS						
	17	18	19	20	21	22	23
A. ITEM RELATED							
1. Reads item	12	10	11	-	-	12	4
2. Read and search	40	40	36	13	14	15	31
3. Generate - trigger	-	-	-	5	-	7	1
4. Generate - no reason	-	-	-	16	-	1	5
5. Generate - reason	-	-	4	6	24	17	3
B. ALTERNATIVES							
1. No comment	14	17	21	15	12	25	8
2. No	1	15	13	11	13	27	53
3. Do not know	13	31	9	-	10	5	-
4. Scans	-	8	4	38	8	2	18
5. Guesses	-	-	-	-	-	-	-
6. Ignores	-	1	15	22	1	5	26
7. Delimiting	-	-	1	-	7	-	-
8. Predisposing	-	-	1	-	1	-	-
9. Oppositions	-	1	-	10	-	-	11
10. Similarities	8	-	-	2	-	68	1
11. Category	-	-	-	-	8	-	-
12. Priority	49	42	9	27	-	1	23
13. Likelihood	1	5	17	19	9	2	5
14. Answer	-	-	-	7	-	1	-
15. Description	-	4	21	-	16	8	-
16. Association	74	36	49	9	75	16	15
C. PROBLEM SOLVING							
1. Restates	15	4	4	2	12	10	1
2. Redescribes	11	36	36	1	33	20	2
3. Generates answer-stem	-	-	4	27	24	25	9
4. Activates hypothesis	12	35	37	-	-	-	-

Number of moves (Continued) - Items 24 - 30

ACTIVITY	ITEMS						
	24	25	26	27	28	29	30
A. ITEM RELATED							
1. Reads item	-	12	15	2	11	9	7
2. Read and search	22	37	39	14	31	34	36
3. Generate - trigger	16	3	1	10	7	5	-
4. Generate - no reason	-	-	-	2	-	-	-
5. Generate - reason	2	-	-	14	2	1	4
B. ALTERNATIVES							
1. No comment	10	18	16	6	21	14	8
2. No	31	30	51	18	33	20	62
3. Do not know	7	27	16	12	29	13	1
4. Scans	-	3	-	12	-	-	-
5. Guesses	-	-	-	-	-	-	-
6. Ignores	-	1	-	-	3	8	-
7. Delimiting	-	4	-	12	-	-	-
8. Predisposing	-	-	-	-	-	-	-
9. Oppositions	4	-	8	-	-	49	1
10. Similarities	1	-	-	-	-	-	2
11. Category	-	-	13	9	4	-	-
12. Priority	-	-	7	-	12	-	67
13. Likelihood	8	5	27	21	5	47	-
14. Answer	-	-	-	-	-	-	-
15. Description	5	10	4	30	1	1	5
16. Association	94	62	18	40	52	8	12
C. PROBLEM SOLVING							
1. Restates	2	12	2	21	8	9	5
2. Redescribes	24	35	9	23	8	-	-
3. Generates answer-stem	8	3	1	26	9	6	4
4. Activates hypothesis	-	-	-	-	-	-	-

Appendix VIII

Frequency Counts - Activities by Item Groups

Number of moves for Group One - Few or No Answers Generated; No Hypotheses Activated

ACTIVITY	ITEMS						
	2	8	10	25	26	29	30
A. ITEM RELATED							
1. Reads item	15	20	3	12	15	9	7
2. Read and search	39	39	38	37	39	34	36
3. Generate - trigger	-	-	-	3	1	5	-
4. Generate - no reason	1	-	2	-	-	-	-
5. Generate - reason	-	-	-	-	-	1	4
B. ALTERNATIVES							
1. No comment	32	7	11	18	16	14	8
2. No	29	1	5	30	51	20	62
3. Do not know	4	1	-	27	16	13	1
4. Scans	-	-	3	3	-	-	-
5. Guesses	4	2	4	-	-	-	-
6. Ignores	-	4	7	1	-	8	-
7. Delimiting	-	1	-	4	-	-	-
8. Predisposing	1	-	-	-	-	-	-
9. Oppositions	2	140	-	-	8	49	1
10. Similarities	-	-	12	-	-	-	2
11. Category	20	-	-	-	13	-	-
12. Priority	-	-	-	-	7	-	67
13. Likelihood	8	2	2	5	27	47	-
14. Answer	-	-	14	-	-	-	-
15. Description	17	-	106	10	4	1	5
16. Association	47	-	-	62	18	8	12
C. PROBLEM SOLVING							
1. Restates	23	8	4	12	2	9	5
2. Redescribes	19	-	18	35	9	-	-
3. Generates answer-stem	1	-	2	3	1	6	4
4. Activates hypothesis	-	-	-	-	-	-	-

Number of moves for Group Two - Few or No Answers Generated; Hypotheses Activated

ACTIVITY	ITEMS 3	7	12	13	17	18	19
A. ITEM RELATED							
1. Reads item	1	9	6	7	12	10	11
2. Read and search	33	37	36	36	40	40	36
3. Generate - trigger	6	2	1	2	-	-	-
4. Generate - no reason	-	-	-	-	-	-	-
5. Generate - reason	1	1	3	2	-	-	4
B. ALTERNATIVES							
1. No comment	21	12	11	17	14	17	21
2. No	7	7	20	14	1	15	13
3. Do not know	8	-	5	1	13	31	9
4. Scans	-	-	-	4	-	8	4
5. Guesses	-	-	-	-	-	-	-
6. Ignores	1	1	5	3	-	1	15
7. Delimiting	-	-	-	3	-	-	1
8. Predisposing	-	-	-	-	-	-	1
9. Oppositions	5	2	-	-	-	1	-
10. Similarities	-	2	-	6	8	-	-
11. Category	1	5	-	3	-	-	-
12. Priority	-	84	-	76	49	42	9
13. Likelihood	9	5	39	6	1	5	17
14. Answer	-	-	-	2	-	-	-
15. Description	47	6	22	6	-	4	21
16. Association	61	36	58	19	74	36	49
C. PROBLEM SOLVING							
1. Restates	30	17	14	12	15	4	4
2. Redescribes	32	25	33	35	11	36	36
3. Generates answer-stem	7	3	4	4	-	-	4
4. Activates hypothesis	39	33	38	38	12	35	37

Number of moves for Group Three - Answers Generated; Few or No Hypotheses Activated

ACTIVITY	ITEMS							
	1	4	5	9	11	14	15	16
A. ITEM RELATED								
1. Reads item	4	3	-	14	7	7	7	6
2. Read and search	8	27	14	31	20	19	19	21
3. Generate - trigger	7	1	3	9	-	13	3	7
4. Generate - no reason	5	10	7	-	17	1	4	2
5. Generate - reason	20	2	15	-	3	5	14	10
B. ALTERNATIVES								
1. No comment	12	19	9	12	18	10	22	9
2. No	11	6	8	19	11	9	17	13
3. Do not know	3	6	3	7	1	1	25	5
4. Scans	4	-	-	-	11	-	2	15
5. Guesses	-	4	-	1	-	-	-	-
6. Ignores	2	65	13	-	73	6	1	1
7. Delimiting	-	-	16	-	-	10	5	-
8. Predisposing	2	-	1	-	-	-	4	1
9. Oppositions	4	-	2	4	2	-	35	-
10. Similarities	-	-	-	11	11	-	-	-
11. Category	6	35	1	13	12	1	-	13
12. Priority	-	-	-	-	9	-	-	46
13. Likelihood	39	18	9	52	7	20	12	13
14. Answer	10	-	10	-	1 -	-	-	8
15. Description	24	6	37	8	-	51	22	4
16. Association	43	5	47	34	4	52	15	32
C. PROBLEM SOLVING								
1. Restates	32	5	14	8	-	5	3	4
2. Redescribes	26	8	10	16	-	34	3	36
3. Generates answer-stem	32	13	25	9	20	19	21	19
4. Activates hypothesis	-	-	-	-	-	-	-	-

Number of moves for Group Three (Continued) - Answers Generated; Few or No Hypotheses Activated

ACTIVITY	ITEMS						
	20	21	22	23	24	27	28
A. ITEM RELATED							
1. Reads item	-	-	12	4	-	2	11
2. Read and search	13	14	15	31	22	14	31
3. Generate - trigger	5	-	7	1	16	10	7
4. Generate - no reason	16	-	1	5	-	2	-
5. Generate - reason	6	24	17	3	2	14	2
B. ALTERNATIVES							
1. No comment	15	12	25	8	10	6	21
2. No	11	13	27	53	31	18	33
3. Do not know	-	10	5	-	7	12	29
4. Scans	38	8	2	18	-	12	-
5. Guesses	-	-	-	-	-	-	-
6. Ignores	22	1	5	26	-	-	3
7. Delimiting	-	7	-	-	-	12	-
8. Predisposing	-	1	-	-	-	-	-
9. Oppositions	10	-	-	11	4	-	-
10. Similarities	2	-	68	1	1	-	-
11. Category	-	8	-	-	-	9	4
12. Priority	27	-	1	23	-	-	12
13. Likelihood	19	9	2	5	8	21	5
14. Answer	7	-	1	-	-	-	-
15. Description	-	16	8	-	5	30	1
16. Association	9	75	16	15	94	40	52
C. PROBLEM SOLVING							
1. Restates	2	12	10	1	2	21	8
2. Redescribes	1	33	20	2	24	23	8
3. Generates answer-stem	27	24	25	9	18	26	9
4. Activates hypothesis	-	-	-	-	-	-	-